biosciences
eastern and central africa

Biosciences eastern and central Africa (BecA),
Nairobi, Kenya.

Introduction to Bioinformatics
May 2008

Etienne de Villiers (PhD)

http://hpc.ilri.cgiar.org/TDR2008/

Adapted from a course originally developed by Dr. David Lynn
Molecular Population Genetics Lab.,
Department of Genetics,
Trinity College Dublin
Ireland

Acknowledgements

This course was adapted from a course designed and implemented by David Lynn and Andrew Lloyd while working at the Education and Research Centre (ERC) at St. Vincent's University Hospital, Dublin. The original course and manual implemented by David Lynn grew naturally from The ABC Bioinformatics Course, an earlier Irish National Centre for BioInformatics (INCBI) project based on GCG and the WWW, to which Aoife McLysaght (TCD) was a major contributor. That in turn owes a debt of gratitude to the ABCT tutorial designed by Rodrigo Lopez when he was the Norwegian EMBnet node. This course would never have got off the ground without the encouragement of Cliona O'Farrelly, the Research Director at the Education and Research Centre (ERC) at St. Vincent's University Hospital. The development of the original course was funded by the Dublin Molecular Medicine Centre and the Conway Institute, University College Dublin.

Table of Contents

INTRODUCTION TO BIOINFORMATICS **4**

DATABASES **6**

SEQUENCE FORMATS **11**

ACCESSION NUMBERS **11**

INTERROGATING (SEQUENCE) DATABASES **14**

SRS - HTTP://SRS.EBI.AC.UK/ 14
ENTREZ - HTTP://WWW.NCBI.NLM.NIH.GOV/ENTREZ/ 19

NUCLEIC ACID SEQUENCE ANALYSIS **20**

1) TRANSLATING DNA IN 6-FRAMES: 20
2) REVERSE COMPLEMENT & OTHER TOOLS: 20
3) OLIGO CALCULATOR - HTTP://WWW.PITT.EDU/~RSUP/OLIGOCALC.HTML 22
4) PRIMER DESIGN 22

PROTEIN SEQUENCE ANALYSIS **25**

1) PHYSICO-CHEMICAL PROPERTIES: 25
2) CELLULAR LOCALIZATION: 27
3) SIGNAL PEPTIDES: 28
4) TRANSMEMBRANE DOMAINS: 30
5) POST-TRANSLATIONAL MODIFICATIONS: 32
6) MOTIFS AND DOMAINS 33
7) SECONDARY STRUCTURE PREDICTION 34

ACCESSING COMPLETED GENOMES **37**

GENEDB - HTTP://WWW.GENEDB.ORG/ 37
TIGRDB - HTTP://WWW.TIGR.ORG/DB.SHTML 41
GENE INDEX PROJECT 43
AND SEVERAL OTHER. 45
ENSEMBL - HTTP://WWW.ENSEMBL.ORG/ 45
NCBI - HTTP://WWW.NCBI.NLM.NIH.GOV/GENOMES/INDEX.HTML 47
ACCESSING THE OTHER GENOMES: HTTP://WWW.NCBI.NLM.NIH.GOV/GENOMES/INDEX.HTML 50

HOMOLOGY SEARCHING **54**

INTRODUCTION 54
BLAST: HTTP://WWW.NCBI.NLM.NIH.GOV/BLAST/ 54
FASTA: 55
SMITH-WATERMAN: 55
OPTIONS IN BLAST. 56
WWW ACCESS TO BLAST. 59
ILRI-BECA BLAST SITE 59
BLAST GUIDELINES. 59

MULTIPLE SEQUENCE ALIGNMENT **63**

CLUSTALW 64
T-COFFEE 66
MULTIPLE SEQUENCE ALIGNMENT EDITORS 67

PRINTED SOURCES ABOUT BIOINFORMATICS & THE INTERNET. **68**

APPENDIX I **69**

APPENDIX II **71**

Introduction to Bioinformatics

This course is designed to impress upon you that computers and the Internet can not only make your work as a biologist easier and more productive but also enable you to answer questions that would be impossible without computational help. Thus there are some computational analyses that you could conceivably do on the back of an envelope or with a pocket calculator and there are others so computationally demanding that you would not attempt them without electronic help. An example of the first would be to scan the following DNA sequence for ecoRI restriction endonuclease sites (GAATTC):

>Adhr *D.melanogaster*
ATGTTCGATTTGACGGGCAAGCATGTCTGCTATGTGGCGGATTGCGGAGGGAGACCAGC
AAGGTTCTCATGACCAAGAATATAGCGAAACTGGCCATTCGGAAAATCCCCAGGCCATC
GCTCAGTTGCAGTCGATAAAGCCGAGTACTTCTGGACCTACGACGTGACCATGGCAA**GA**
ATTCATATGAAGAAGTACTGATGGTCCAAATGGACTACATCGATGTCCTGATCAATGGT
GCTACGCTGATAACATTGATGCCACCATCAATACAAATCTAACGGGAATGATGAACACG
TGTTACCCTATATGGACAGAAAAATAGGAGGAATTCGTGGGCTTATTGTTCGGTCATTG
GATTGGACCCTTCGCCGGTTTTCTGCGCATATAGTGCAGTGTAATTGGATTTACCAGAA
GTCTAGCGGACCCTCTTTACTATTCCCAGCTGTGATGGCGGTTTGTTGTGGTCCTACAA
GGGTCTTTGTGGACCGGGGTTTTTAGAATACGGACAATCCTTTGCCGATCGCCTGCGGC
GAGCGCCCCATCGGTTTGTGGTCAGAATATTGTCAATGCCATCGAGAGATCGGAGAATG
GATTGCGGATAAGGGTGGACTCGAGTTGGTCAAATTGCATTGGTACTCGACCAGTTCGT
GCACTATATGCAGAGCAATGATGAAGAGGATCAAGAT

(This sequence is written in Fasta format - see below for sequence formats.)

A computer could do it quicker, but it is still trivial to do it by eye. Especially as one of the sites has been picked out in bold. Can you find the other(s)? Sequence analyses impossible without a computer include, but are not limited to, most operations that involve the sequence databases. The DNA databases (Genbank EMBL DDBJ) are curated by three different groups in Bethesda, MD, Hinxton, UK and Mishima, JP but, because they exchange information on a daily basis, should be effectively the same in content. The DNA databases are doubling in size about every year; they currently (Oct 2005) comprise:

> 100 gigabases of sequence.

So finding all of the ecoRI sites in GenBank or even the whole of a printed copy of the human genome (3,200,000,000 bp) would take more than a few minutes.

This course will introduce you to some of the more commonly used bioinformatics tools; tell you how to use them and, more importantly, how to use them "correctly" or at least more effectively. Most of the analysis will be carried out on the World Wide Web (WWW). This is partly because it is available to all comers without requiring direct access to the necessary computers, which serve as database and software repositories. But it is also partly because a well-designed Web site can be particularly user-friendly and intuitive in its operations.

There are likely to be network related problems trying to make 25 simultaneous connections over the Internet to the same site. Try doing the course exercises late in the evening, early in the morning (best for speed!) or at weekends.

This module in bioinformatics is designed to give you a flavour of what analytical and informative tools are available on the World Wide Web.

Bioinformatics

Bioinformatics has been described as the storage, retrieval and analysis of biological sequence information. In this short course we will be taking a broader definition: how computers can maximise the biological information available to you. This will touch on determining the 3-D structure of bio-molecules and trying to relate this to their function as well as accessing the relevant literature. I hope that, by the end of the course, everyone will be adopting a more explicitly evolutionary understanding of 'their' molecule. The formal course practicals can be carried out entirely on the World Wide Web using Netscape or the other Web-browser. Nevertheless, we recommend using locally installed (FREE) software for the phylogenetic trees part of the course.

You should note that several important types of bioinformatic analysis are not freely accessible on the Web, but are available on various password controlled computers. In particular, types of analysis that require large amounts of computational power/time are best carried out off the web. Analyses of many genes are also often better done in an environment where a computer program does the pointing and clicking for you. For the record, the GCG package is a suite of programs which carry out almost all the analyses that a molecular biologist might want to do with/on DNA or protein sequences (secondary structure prediction, two sequence alignment, conceptual translation of DNA, restriction site analysis, primer design, as well as homology searching, multiple sequence alignment etc.). For phylogenetic inference and tree drawing, the PHYLIP package (versions available for PCs, Macs and Unix) will answer most needs. Both of these software packages and a variety of other sequence analysis packages are available on the Internet.

The web, by contrast, is a total mess: the same program is implemented with different defaults at different sites; it is often not clear what those defaults, options and parameters are; the results are not easily transferred to a different program. So it is free, but there is a cost! You are advised to validate any analysis against the results yielded by other sites.

Databases

Databases are of course the core resource for bioinformatics. There is plenty of software for analysing one or a few sequences, but many of the computationally interesting and biologically informative programs access databases of information. Frequently used classes are the biological sequence databases. These include:

- EMBL (European Mol Biol Lab)
- GenBank
- DDBJ (DNA DB of Japan)

These three DNA databases exchange their data on a daily basis and so should be identical as to content. They are, however, rather different in format:
Each of the database cited above consists of a (very large number) of entries, each consisting of a single sequence preceded by a quantity of 'annotation' that puts the sequence in its biological, functional and historical context. Without the annotation, GenBank would be a meaningless string of 32 billion As Ts Cs and Gs. Compare and contrast the two extracts from a) EMBL and b) Genbank (DDBJ has the same look-and-feel as Genbank):

a) EMBL

```
ID ECRECA standard; DNA; PRO; 1391 BP.
AC V00328; J01672;
DT 09-JUN-1982 (Rel. 01, Created)
DT 12-SEP-1993 (Rel. 36, Last updated, Version 4)
DE E. coli recA gene.
KW .
OS Escherichia coli
OC Bacteria; Proteobacteria; gamma subdiv; Enterobacteriaceae;
OC Escherichia.
RN [1]
RP 1-1374
RX MEDLINE; 80234673.
RA Sancar A., Stachelek C., Konigsberg W., Rupp W.D.;
RT "Sequences of the recA gene and protein";
RL Proc. Natl. Acad. Sci. U.S.A. 77:2611-2615(1980).
```

b) GenBank

```
LOCUS ECRECA 1391 bp DNA BCT 12-SEP-1993
DEFINITION E. coli recA gene.
ACCESSION V00328 J01672
KEYWORDS .
SOURCE Escherichia coli.
ORGANISM Escherichia coli
Eubacteria; Proteobacteria; gamma subdiv; Enterobacteriaceae;
Escherichia.
REFERENCE 1 (bases 1 to 1374)
AUTHORS Sancar,A., Stachelek,C., Konigsberg,W. and Rupp,W.D.
TITLE Sequences of the recA gene and protein
JOURNAL Proc. Natl. Acad. Sci. U.S.A. 77 (5), 2611-2615 (1980)
```

You can see that these two are obviously talking about the same sequence from *E.coli*, but the information is encoded in a rather different way. This makes no difference to us reading the text, but causes problems when writing a program to interrogate a database.

Each database entry has a name, called ID or LOCUS, which tries to be mnemonic and marginally informative. More importantly each has an accession number which is arbitrary but which remains attached to the sequence for the rest of time. The organism might become reclassified, the gene may get renamed and the ID is thus subject to change, but by noting the accession number you should always be able to identify and retrieve the sequence. Note also that the original publication is cited. Usually there will be other papers documenting functional analysis, mutations, allelic variations, 3-D structure and so on.

Further down in the entry is annotation about the sequence itself, so that the sequence is parsed into meaningful bits called a features table:

a) EMBL

```
FT source 1. .1391
FT /organism="Escherichia coli"
FT /db_xref="taxon:562"
FT mRNA 191. .>1391
FT /note="messenger RNA"
FT RBS 229. .233
FT /note="ribosomal binding site"
FT CDS 239. .1300
FT /db_xref="SWISS-PROT:P03017"
FT /transl_table=11
FT /gene="recA"
FT /product="recA gene product"
FT /protein_id="CAA23618.1"
FT mutation 353. .353
```

b) GenBank

```
FEATURES Location/Qualifiers
source 1..1391
/organism="Escherichia coli"
/db_xref="taxon:562"
mRNA 191..>1391
/note="messenger RNA"
RBS 229..233
/note="ribosomal binding site"
gene 239..1300
/gene="recA"
CDS 239..1300
/gene="recA"
/codon_start=1
/transl_table=11
/product="recA gene product"
/db_xref="SWISS-PROT:P03017"
mutation 353
/gene="recA"
```

Again you can see that the information exchange between Genbank and EMBL includes all significant portions of the annotation. Such useful signals and data as the open reading frame (CDS for CoDing Sequence), the ribosome binding site, intron boundaries, signal peptides, variants/mutations may be recorded.

Protein databases:

- SwissProt
- PIR (Protein Information Resource)
- GenPept

a) Swissprot

```
ID RECA_ECOLI STANDARD; PRT; 352 AA.
AC P03017; P26347; P78213;
DT 21-JUL-1986 (REL. 01, CREATED)
DT 21-JUL-1986 (REL. 01, LAST SEQUENCE UPDATE)
DT 15-DEC-1998 (REL. 37, LAST ANNOTATION UPDATE)
DE RECA PROTEIN.
GN RECA OR LEXB OR UMUB OR RECH OR RNMB OR TIF OR ZAB.
OS ESCHERICHIA COLI, AND SHIGELLA FLEXNERI.
OC BACTERIA; PROTEOBACTERIA; GAMMA SUBDIVISION; ENTEROBACTERIACEAE;
OC ESCHERICHIA.
...
...
CC -!- FUNCTION: RECA PROTEIN CAN CATALYZE THE HYDROLYSIS OF ATP IN THE
CC PRESENCE OF SINGLE-STRANDED DNA, THE ATP-DEPENDENT UPTAKE OF
CC SINGLE-STRANDED DNA BY DUPLEX DNA, AND THE ATP-DEPENDENT
CC HYBRIDIZATION OF HOMOLOGOUS SINGLE-STRANDED DNAS. IT INTERACTS
CC WITH LEXA CAUSING ITS ACTIVATION AND LEADING TO ITS AUTOCATALYTIC
CC CLEAVAGE.
CC -!- INDUCTION: IN RESPONSE TO LOW TEMPERATURE. SENSITIVE TO
CC TEMPERATURE THROUGH CHANGES IN THE LINKING NUMBER OF THE DNA.
CC -!- DATABASE: NAME=E.coli recA Web page;
CC WWW="http://monera.ncl.ac.uk:80/protein/final/reca.htm".
KW DNA DAMAGE; DNA RECOMBINATION; SOS RESPONSE; ATP-BINDING; DNA-
BINDING;
KW 3D-STRUCTURE.
FT INIT_MET 0 0
FT NP_BIND 66 73 ATP.
FT CONFLICT 112 112 D -> E (IN REF. 5).
FT TURN 4 4
FT HELIX 5 21
FT HELIX 23 25
FT TURN 29 30
```

etc etc

b) PIR

```
>P1;RQECA
recA protein - Escherichia coli
C;Species: Escherichia coli
C;Date: 31-Jul-1980 #sequence_revision 14-Nov-1997 #text_change 14-Nov-1997
C;Accession: G65049; A93847; A93846; S11931; S63525; S63979; A03548
...
C;Comment: The recA protein plays an essential role in homologous recombination, in
induction of the SOS response, and in initiation of stable DNA replication.
C;Genetics:
A;Gene: recA
```

```
A;Map position: 58 min
C;Superfamily: recA protein
C;Keywords: ATP; DNA binding; DNA recombination; DNA repair; P-loop; SOS response
F;67-75/Region: nucleotide-binding motif A (P-loop)
F;141-145/Region: nucleotide-binding motif B
F;73/Binding site: ATP (Lys) #status predicted
```

Note that these two entries refer to the same gene from E.coli despite differences in the way the data is encoded. However, in contrast to the difference between EMBL and Genbank, the quality of the annotation is quite different. The 3-D structure of this gene has been worked out and this information is reflected in the SwissProt entry as the position of every alpha-helix and beta-sheet is noted. In general, the quality of the annotation and the minimization of internal redundancy makes SwissProt the preferred database to use. However, note that PIR records the Genetic Map position of the gene; so it is probably good to scrutinize both databases to abstract maximal information.
SwissProt also gives added value by incorporating a large number of DR (database reference) tags, pointing to equivalent information in other databases.

a) SwissProt:

```
DR EMBL; V00328; G42673; -.
DR EMBL; X55553; -; NOT_ANNOTATED_CDS.
DR EMBL; AE000354; G1789051; -.
DR EMBL; D90892; G1800085; -.
DR PIR; A03548; RQECA.
DR PIR; S11931; S11931.
DR PDB; 1REA; 31-OCT-93.
DR PDB; 2REB; 31-OCT-93.
DR PDB; 2REC; 01-APR-97.
DR PDB; 1AA3; 23-JUL-97.
DR SWISS-2DPAGE; P03017; COLI.
DR ECO2DBASE; C039.3; 6TH EDITION.
DR ECOGENE; EG10823; RECA.
DR PROSITE; PS00321; RECA; 1.
DR PFAM; PF00154; recA; 1.
```

When these are used as hypertext links they can enable a WWW browser to locate an extraordinary depth of detail about a given entry, 3-D structure (PDB), protein motifs (Prosite), families of related genes (Pfam), the DNA sequence (EMBL) and a couple of specialist E.coli added-value databases. SRS is one program that makes these hypertext links. The PIR cross-references are far fewer and less explicit; its reference to Genbank (GB:U00096) refers to the whole E.coli genome, whereas SwissProt points specifically to the gene (DR EMBL; V00328)

b) PIR

```
...
A;Cross-references: GB:AE000354; GB:U00096; NID:g2367149; PID:g1789051; UWGP:b2699
```

All these databases are made up of entries, concatenated one after the other in plain readable text. As such they are far bigger than necessary if you are trying to analyze the sequence rather than interrogate or browse the annotation. For these purposes, special high-compressed databases can be constructed. Frequently these are not readable by humans because they have been optimized for speed reading computers.

One of the simplest compression protocols is called Fasta format in which the annotation is edited down to a single title line followed by the sequence. The sequence at the top of the chapter is in Fasta format. All protein databases use the one-letter amino acid code, can you think why this might be?

Sequence Related Databases

Not all biologically relevant Databases consist of sequences and annotation. There are databases of journal abstracts, taxonomy, 3-D structures, mutations and metabolic pathways. Some of the most useful of these are databases which specialise in particular entities that can be found dispersed in the "whole sequence" databases.

You notice one of the cross-references for the SwissProt entry is:

DR PROSITE; PS00321; RECA; 1.

Prosite is a database of protein motifs. PS00321 is a family of proteins that all have the motif:
PA A-L-K-F-[FY]-[STA]-[STAD]-[VM]-R

and are all believed to bind DNA, hydrolyze ATP and act as a recombinase. One of the members of this family is the recA gene in E.coli which gives its name to PS00321. In the pattern above, the residues within [square brackets] are alternatives. Convince yourself that ALKFFAAVR could belong to the family but ALKFAAAVR could not. There are more than 1000 other families classified in a similar way. Finding a Prosite link in a SwissProt gene is a great help in finding other proteins related by structure and/or function.

Interpro - http://www.ebi.ac.uk/interpro/

You should also be aware of the Interpro project which incorporates and sorts data from a diversity of protein motif and domain databases into one searchable meta-database.

Sequence formats

As we have seen comparing database entries above, there are dozens of different ways in which you can store or represent the same fundamental information. Databases are often compiled in, highly conventionalized, readable English text. Computers, being not so bright, will have difficulty reading and interpreting the information unless the conventions are quite rigidly obeyed. There are a very large number of ways you can write, store and transmit simple one-dimensional sequence files. A common sequence interchange program called 'readseq' recognizes at least 22 different file formats. If a computer program does not recognize the format of an input sequence it may not work or, worse, misinterpret header lines as sequence data or otherwise mangle your analysis. Some commonly used file/sequence formats are shown below:

1) Fasta (named for a widely used homology searching program) – single title line beginning >:

```
>ECRGCG TRANSLATE of: ecrgcg 1 to: 1062
MAIDENKQKALAAALGQIEK
ALGAGGLPMGRIVEIYGPES
TPKAEIEGE*
```

2**) Staden** (named after Rodger Staden - early, but still extant, software writer) – same as raw sequence:

```
MAIDENKQKALAAALGQIEK
ALGAGGLPMGRIVEIYGPES
TPKAEIEGE*
```

3) NBRF/PIR (named after the protein database):

```
>P1;ecrgcg.pep
ecrgcg.pep, 354 bases, 218 checksum.
MAIDENKQKA LAAALGQIEK
ALGAGGLPMG RIVEIYGPES
TPKAEIEGE*
```

Accession numbers

The information above makes you aware of the diversity of ways in which something so simple as a one-dimensional sequence may be represented. Another source of confusion is the variety of identifying numbers attached to sequences and knowing to which database they refer. Accession numbers are used as unique and unchanging numbers. They are not mnemonic, although databases also have a less stable, more memorable nomenclature: HBB_HUMAN, HSHBB, HUMHBB 2HBB are all human beta globin IDs in various databases,

- **GenBank/EMBL** accession numbers: originally a letter followed by 5 digits (X32152, M22239). When the number of sequences exceeded 2,600,000 - 2 letters followed by 6 digits (AL234556, BF345788).
- **SwissProt**. Still one letter followed by 5 digits, letter is either O,P,Q. P23445.
- **PIR**: the 'other' protein database, one letter followed by 5 digits, but numbers confusable with EMBL/GenBank: B93303 is chimp haemoglobin in PIR but a random genomic clone fragment in EMBL.
- **GenPept**. Conceptual translations from DNA that have not yet been annotated well enough to get into SwissProt. three letters and five digits, e.g.: AAA12345.
- **Trembl** (Translated EMBL): O, P or Q followed by 5 letters/digits.
- **PDB** protein structure records: 1 digit and three letters 1HBA, 1TUP

More recently, an attempt has been made to reduce the redundancy in the databases (there were 180 copies of D. melanogaster alcohol dehydrogenase each with its own accession number). One result is RefSeq - NCBI's "reference sequence" database

RefSeq: Two letters, and underscore bar, and six digits,
mRNA records (NM_*) NM_000492 **genomic DNA contigs** (NT_*) NT_000347 **curated/annotated Genomic regions** (NG_*) NG_000567 **Protein sequence records** (NP_*) NP_000483

We will see how RefSeq is becoming the central resource for gene characterization, expression studies, and polymorphism discovery. Because of the high level of necessary curation, it is not anywhere close to being comprehensive even for those species that are included.

Accession numbers give the community a unique label to attach to a biological entity, so we all know we are talking about the same thing. Sequences in databases evolve as their real biological counterparts do. They need to be updated, corrected and merged and we need to know which version of the sequence entry is being referred to. GenBank has used gi numbers and, more recently, version numbers for this. Each small change made to a Genbank record gets the next gi number e.g. gi6995995 and so is totally arbitrary. Version numbers are appended to the accession number after a dot – V00234.2, NM_000492.2.

The other programs to use in the course are many and varied. We have tried to put links to them all on the course website:

http://hpc.ilri.cgiar.org/TDR2008

A few overall points for the course:

• Take the opportunity to compare and contrast different methods of doing a particular analysis.

• By all means take the defaults but be aware that changing them will almost certainly get more or better information.

• The Web is free, and you get what you pay for, so use the Web with care & caution.

• As with lab work it takes time to get the protocol working. Once you have one that works for you, write it down, bookmark and remember it. But note, the Web changes rapidly and you cannot afford to use outmoded technology for long.

Interrogating (sequence) databases

SRS - http://srs.ebi.ac.uk/

The DNA databases are enormously rich information resources partly because they are so big, but it would make little sense if it consisted of a long list of As Ts Cs and Gs. There are millions of individual entries in EMBL. An entry could be a fragment as short as 3 base pairs (e.g. M23994) or a large contig consisting of many genes, including complete eukaryotic chromosomes (e.g. X59720). The value of the database lies substantially in the quality of the annotation which puts the sequence in its biological context.
As a biologist you may need to be able to interrogate the Database to find particular sequences or a set of sequences matching given criteria, such as:

The sequence published in Cell 31: 375-382
All sequences from Aspergillus nidulans
Sequences submitted by Peter Arctander
Flagellin or fibrinogen sequences
The glutamine synthase gene from Haemophilus influenzae
The upstream control region of Bacillus subtilis Spo0A

SRS (Sequence Retrieval System) is a very powerful, WWW-based tool, developed by Thure Etzold at EMBL and subsequently managed by Lion Biosciences, for interrogating databases and abstracting information from them.

One of the neatest features of SRS is the fact that interrelated databases can be cross-referenced with WWW hypertext links. This means that you can discover the protein sequence, the cognate DNA sequence, a family of related proteins in other species, a Medline reference to read an abstract of the original publication, a 3-D structure - all with a few point-and-clicks with the mouse.

There are several SRS servers on the Web. We will be using

http://srs.ebi.ac.uk/

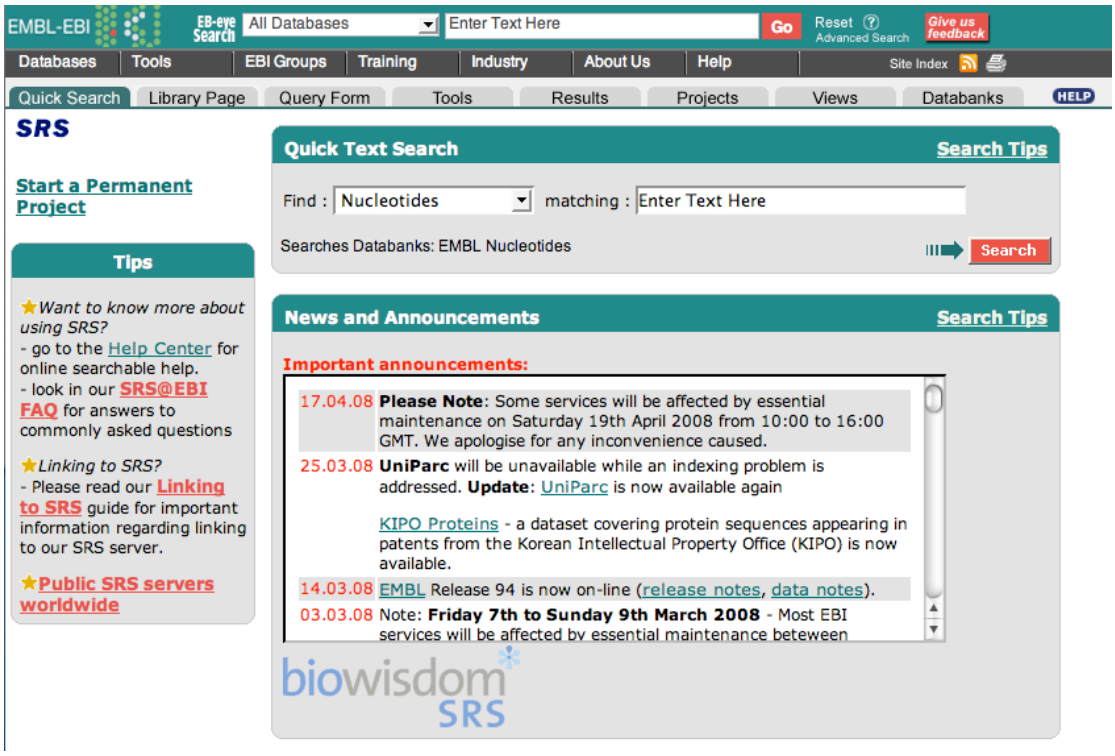
at the EBI in England because a) it has a large number of interlinked databases b) connectivity to the UK is good c) they are attempting to interconnect their SRS server with their clustalW server and blast server. If the SRS server at the EBI is slow you might try :

http://srs.hgmp.mrc.ac.uk/
http://srs.sanger.ac.uk/

The three servers (EBI, HGMP and Sanger) are all located within a few metres of each other on the Wellcome Trust Genome Campus at Hinxton in England.

The documentation for SRS is getting better. With experience and practice you will get to use as much of SRS's power as necessary to obtain the results you need. I will show below, as a worked example, a series of instructions to obtain the sequences of all the mammalian osteonectin proteins in SwissProt, and download them locally to carry out a multiple sequence alignment using, say, clustalW. It should also be possible to do the multiple alignment on the EBI clustalW server.

Use your browser (Netscape?) to go to http://srs.ebi.ac.uk/ or one of the other SRS servers at the top of the Course page. You should see the following screen with several tabs on top:


EMBL-EBI
EB-eye Search
All Databases
Enter Text Here
Go
Reset
Advanced Search
Give us feedback
Databases
Tools
EBI Groups
Training
Industry
About Us
Help
Site Index
Quick Search
Library Page
Query Form
Tools
Results
Projects
Views
Databanks
HELP
SRS
Start a Permanent Project
Tips
Want to know more about using SRS?
- go to the Help Center for online searchable help.
- look in our SRS@EBI FAQ for answers to commonly asked questions
Linking to SRS?
- Please read our Linking to SRS guide for important information regarding linking to our SRS server.
Public SRS servers worldwide
Quick Text Search
Search Tips
Find : Nucleotides
matching : Enter Text Here
Searches Databanks: EMBL Nucleotides
Search
News and Announcements
Search Tips
Important announcements:
17.04.08 Please Note: Some services will be affected by essential maintenance on Saturday 19th April 2008 from 10:00 to 16:00 GMT. We apologise for any inconvenience caused.
25.03.08 UniParc will be unavailable while an indexing problem is addressed. Update: UniParc is now available again
KIPO Proteins - a dataset covering protein sequences appearing in patents from the Korean Intellectual Property Office (KIPO) is now available.
14.03.08 EMBL Release 94 is now on-line (release notes, data notes).
03.03.08 Note: Friday 7th to Sunday 9th March 2008 - Most EBI services will be affected by essential maintenance beteween
biowisdom SRS

Click on tab '**Library Page**'.

This takes you to what is called the TOP PAGE. This page allows you to choose the database(s) that you wish to search. The databases may be of various types, including:

Literature, Bibliography and Reference Databases: MEDLINE, Taxonomy

Nucleotide sequence databases: EMBL, Patent DNA, EMBL (Contig)

UniProt Universal Protein Resource: UniprotKB, UniprotKB/Swiss-Prot

For more information about the contents of the database place cursor on the relevant blue underlined hypertext link - UniProtKB say, or click link for more information.

- **Click the box [_] to the left of UniProtKB.**
- **Click on the Query Form tab at the top of the page**

This will move you to a Query Form Page that permits you to submit particular queries (such as have been suggested at the beginning of this chapter) to the databases. At the top of this page will be a note of which database(s) you have chosen to search and a block of four text-insert boxes which you can use to enter your question.
to the left you will see some things you can change including:

1. **[Reset]** - which clears the screen.
2. **combine search terms &(AND)** - which enables you to apply other logical (boolean) operators.
3. **Use wildcards** which means that "bact" will be interpreted as bact* and look for bacteria, bacteriophage, etc.
4. **Results type** to get.
5. **Number of entries to display per page** (default is 30)

Your question can be entered into one of more of the text-insert boxes, thus:

- **Click [All text] change to [Description]** and **insert "serum resistance associated" in box**

Note: it does not have to be "serum resistance associated" it could be ubiquitin or haemoglobin or hemoglobin or actin & alpha. Separate keywords in the same box have to be linked by a logical (Boolean) operator such as
and: &
or: |
but not: !

- **Click the next [All text] change to [Taxonomy] and insert "trypanosoma" in box**
- **Click Search**

a new window appears with Query "([uniprot-Description: serum resistance associated*] & [uniprot-Taxonomy:Trypanosoma*]) " found 4 entries. This is how SRS interprets what you have entered in the boxes and the numbers of "hits" found.

- **Under Display options change [UniprotView] to [FastaSeqs]**
- **Click [Save]**

  Make sure view is [FastaSeqs].
- **Click [Save]**
- **Output To:** select **File (text)**
- **Click [Save]**
- **Change selection .../wgetz to .../serum_resist.pro and then Click Save.**

This should dump the concatenated fasta format protein sequences into a local file called serum_resist.pro. You can use this file as input for **clustalw**. There may be

local security difficulties with downloading sequences onto a public terminal - check with your neighbours or your demonstrator.

Query manager: a powerful tool

A quick example will show how you can combine very complex queries to zero in on the sequence(s) you need.
Having selected your database(s) go to the Query Form Page and enter:

- **[Description] calmodulin**

you should get about 1865 entries.

- **Click [QUERY FORM] tab** at the top of the page to get a new page and enter:
- **[Organism name] human** (or indeed Homo sapiens)

this will get you a large number of sequences.

- **Click [RESULTS] tab** at the top of the page

A new window should appear with the results for all the queries you have entered in the current SRS session. In the top box of this page **enter "Q1 & Q2"** (leave off the quotes!) and press **Search**.

Note: Your mileage may vary here. Q1 and Q2 may refer to earlier queries in this SRS session such as for serum resistance associated proteins in Trypanosomes so use good judgement.

You have just used a boolean logical expression to yield sequences which are a) human and b) have "calmodulin" in the SwissProt description. This shows you how it can be unreliable to depend on the annotation to get homologous sequences. Nevertheless, the list should contain the SwissProt entry for **CALM_HUMAN** which is what you want.

Questions

1. Can you think of a better way to find other mammalian calmodulin genes?

2. If you do a search in SwissProt for "calmodulin" using the [AllText] descriptor instead of [Description] you find many more entries, why do you think you get more entries under this search?

3. There are more entries in SwissProt under [Organism] dog than [Author] dog, but more for [Author] wolf than [Organism] wolf. Why do you think this is so?

4. Searching [Organism] mouse in SwissProt yields some plant sequences: prove this by finding sequences matching [Organism] mouse & [Taxon] viridiplantae. Why is this so? (Clue: append wildcard *).

You should be able to reveal the full SwissProt entry for any protein sequence. If you do this you will see several (? blue, underlined) hypertext links to related databases. Almost certainly at least one of these will be EMBL and one to Medline. Probably one will be the prosite motif database. If the 3-D structure is known, one link will be to PDB. Investigate these other databases to get as much relevant information as possible about your sequence.

Aside: Displaying 3-D structures is not "fitted as standard" on all terminals. You may need to get a copy of the RasMol 3-D structure viewer and install it in such a way that your Netscape/IE will recognise it and connect suitable (3-D sequence) file to it.

To display a PDB entry of 3-D coordinates as a rotatable, colorable model you need to click on the **[save]** button. The change the "use mime type" choice-box to **chemical/x-pdb** and then click on the **[save]** box. This should fire up CHIME a WWW implementation of RasMol.) Your mileage may vary!

It is this, interlinked databases, aspect of SRS which gives it a large part of its power. You can extend your search to include other sequences related in some particular (or peculiar!) way. The Prosite link allows you to find members of a protein family. The EMBL link allows you to find the introns and the intron splice junctions, not to mention the ribosome-binding site, the stop codon and the journal reference for the original sequence. The Medline link will give you an abstract etc. You will probably find that:

The PubMed server at http://www.ncbi.nlm.nih.gov/Entrez/ is a far better tool for browsing Medline that what is offered with SRS. Especially powerful is its facility for finding [Related entries].

Additional questions:

"Effective researchers know how to find things out"

1. Who submitted the serum amyloid A (SAA) gene sequence for Canis familiaris?
2. What prosite motif defines the recA family of prokaryotic proteins? Which Dublin-based phylogeneticists used multiple-sequence alignment to define this motif?
3. What are the first and last 5 bases in the intron of the yeast actin gene with EMBL accession number V01288?
4. What is the map position of one of the human SAA genes (SwissProt: P02735)? What cross-reference database is most likely to have map position?
5. What mutation at what position causes phenylketonuria (PKU)? (hint: EMBL K03020) but then try SwissProt: P00439.
6. What bases define the ribosome binding site of the Bacteroides fragilis glnA gene? Perhaps start from the E.coli homolog SwissProt: P06711.
7. Why is the name Saarinen associated with life-threatening cardiac arrythmias? (Hint: not because of architectural flaws...try voltage gated potassium channels)
8. Are there more publicly available DNA sequences from Rodents or Prokaryotes? What about protein sequences?

Entrez - http://www.ncbi.nlm.nih.gov/Entrez/

Entrez is the US equivalent of SRS and is available from the NCBI webpage. You will most likely be familiar with Entrez for interrogating Medline, but the same engine can be pointed at DNA and protein databases. It is handy if you are familiar with the Entrez system and you want a sequence whose name or accession number you already know. At the top of the Entrez page change the Search [__] choice box from PubMed to the appropriate sort of database – the available options are listed on the Entrez page. If you want the sequence alone – to paste into some analysis page – change the Display [__] choice box to FASTA then click on [Save] or [Display] depending on whether you want a permanent or transitory copy of you proteins. Entrez has a more complex syntax for less straightforward queries.

Nucleic Acid Sequence Analysis

TOPICS:

1. Translating DNA in 6 frames.
2. Reverse complement & other tools.
3. Calculating some properties of DNA/RNA sequences.
4. Primer design.

1) Translating DNA in 6-frames:

Translate tool - http://www.expasy.ch/tools/dna.html

This tool allows the 6-frame translation of a nucleotide (DNA/RNA) sequence to a protein sequence in order to locate open reading frames in your sequence. Use the serum_resist.dna sequence on the course website http://hpc.ilri.cgiar.org/TDR2008/

- Go to Translate tool URL above.
- Paste your sequence in the box provided & click "TRANSLATE SEQUENCE".
- You can choose 3 options
  - Verbose – puts Met & Stop to highlight start & stop codons.
  - Compact – useful if you want to use output in other programs.
  - Includes nucleotide sequence – nucleotide sequence is above the translation.
- This returns a 6-frame translation of your sequence. You can then choose the correct frame.

See Appendix II for the genetic code

2) Reverse Complement & other tools:

There are many cases where you might want to obtain the reverse complement of a DNA sequence, for example the reverse complement is needed as a negative control when doing a DNA hybridisation experiment.

Search launcher at Baylor College –

http://searchlauncher.bcm.tmc.edu/seq-util/seq-util.html

This tool contains a number of different applications for nucleic acid sequence analysis: For each application you can click on the following [H] [O] [P] [E] = [H]:Help/description; [O]:full Options form; [P]:search Parameters; [E]:Example search. On all the Baylor pages (and everywhere else possible) it is important to

investigate the options [O] to see a) what are the defaults and b) what options seem worth changing. The following programs are available:

Readseq:

Converts nucleic acid/protein sequences between any of 30 different formats. It is often appropriate to convert to FASTA format. A large number of input formats are permitted. See help for details [H].

RepeatMasker:

RepeatMasker is a program that screens DNA sequences for interspersed repeats known to exist in mammalian genomes as well as for low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked (replaced by Ns). On average, over 40% of a human genomic DNA sequence is masked by the program. This is important in primer design so that you do not design a primer that spans a region with repeats. It is also important before doing a homology search as repeats in your sequence may hit other repeats in the genome (although BLAST now does this for you).

Primer Selection -PCR primer selection (See primer design later).

WebCutter- restriction maps using enzymes w/ sites >= 6 bases.

6 Frame Translation - translates a nucleic acid sequence in 6 frames.

Reverse Complement - reverse complements a nucleic acid sequence.

Reverse Sequence - reverses sequence order.

Sequence Chopover - cut a large protein/DNA sequence into smaller ones with certain amounts of overlap.

HBR - Finds E.coli contamination in human sequences.

Exercise: Paste in your own sequence of interest or alternatively examine an example output for each application by clicking [E] beside each program. Pay particular attention to the options available: these will give you clues about standard practice.

3) Oligo Calculator - http://www.pitt.edu/~rsup/OligoCalc.html

Tool to calculate the **length**, **%GC content**, **Melting temperature (Tm)** the midpoint of the temperature range at which the nucleic acid strands separate, **Molecular weight**, & what an **OD** = 1 is in picoMolar of your input nucleic acid sequence.

Many of these parameters are useful in primer design (see next section) and in other areas of molecular biology.

- Go to URL above.
- Paste the Homo sapiens interleukin 11 (IL11) sequence on the course webpage in the box provided & click "Calculate".

Example:

>gi|10834993|ref|NM_000641.1| Homo sapiens interleukin 11 (IL11), mRNA
Length = 2281
% GC content = 55
Tm = 87 °C
Molecular Weight = 704856 daltons (g/M)

OD of 1 = 41 picoMolar

4) Primer design

(Originally written in Jan 2002 by Dr. Norma O'Donovan (Thanks!).)

The recommended site (although there are several others available on the web) is GeneFisher:

http://bibiserv.techfak.uni-bielefeld.de/genefisher/help/wwwgfdoc.html

The submission form:

http://bibiserv.techfak.uni-bielefeld.de/cgi-bin/gf_submit?mode=STARTUP&sample=dna

The input form is straightforward and well documented.

Primer Design Tips:

- **Primer Length:** usually between 18 and 24 base pairs
- **% GC**: Optimum GC content is 45-55%
- **Annealing Temperature:** Should be between 55 ^{o}C and 65 ^{o}C and ideally the annealing temperature of the 2 primers should be similar. A quick equation (Wallace formula) for calculating the annealing temperature of the primer is:

2 x (no. of As + Ts) + 4 x (no. of Gs + Cs)

The lower of the 2 primer annealing temperatures is the highest temperature that can be used for annealing. (Usually when optimising PCR you would start with an annealing temp. a few degrees below the Tm of the primers).

- **G/C clamps:** The 3' end of the primer should be able to form G/C clamps, i.e. several consecutive G/C or C/G base pairs, between the 3' end of the primer and the template DNA.
- **Length of PCR product:** The optimum size is 100 – 500 base pairs for conventional PCR. Shorter products can be used for real time PCR or longer products can be amplified using special polymerases.

Things to avoid:

1. Complimentarity within a primer or between 2 primers (especially in the ends), used in the same reaction, as this may cause primer dimers.

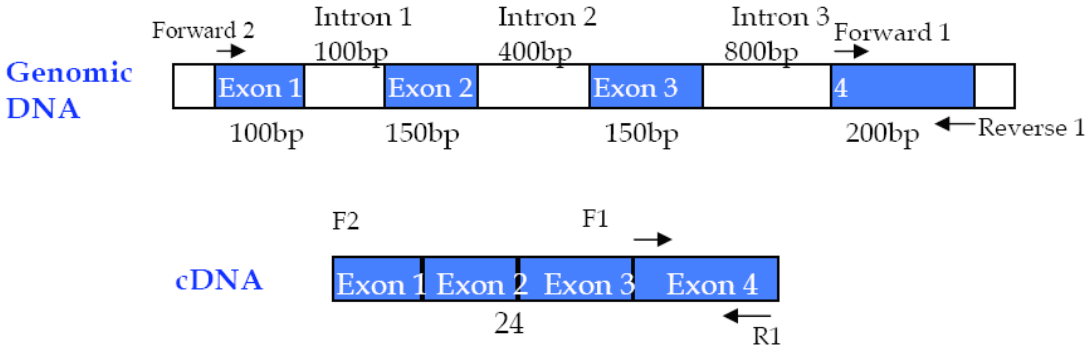
2. Strings of a single nucleotide (more than 3).

3. Non-specific binding of primers to related sequences – check the specificity of the primers by doing a BLAST search of the database (non-redundant and genomic) with each of the primer sequences.

Primers for RT-PCR

The same rules as above apply but there are a few extra considerations. If you are doing RT-PCR with total RNA there may be genomic DNA contamination present in the RNA. (You can DNase treat to remove it or purify poly-A mRNA). If it is not removed you must ensure that your primers specifically amplify the cDNA (complementary to mRNA). Ideally the primers should not amplify the genomic DNA at all but if that is not possible the genomic product should be distinguishable from the cDNA product on a gel, based on size. Therefore the primers must span at least one intron in the genomic DNA. To identify the position of introns in the sequence align the mRNA sequence with the genomic sequence using a pairwise BLAST sequence alignment (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html). Alternatively, for human or mouse sequences, on the UCSC website (http://genome.ucsc.edu/) you can do a BLAT search with the mRNA which will identify the intron/exon structure of the gene.

Example:



If the forward and reverse primers are designed in exon 4 the PCR product obtained from the cDNA will be the same size as the genomic PCR product. If the forward primer is in exon 1 and the reverse primer is in exon 4 the cDNA product will be approx. 600bp whereas the PCR product from genomic DNA would be about 1900bp, which probably wouldn't be amplified in conventional PCR.

Protein Sequence Analysis

TOPICS

1. Physico-chemical properties.
2. Cellular localization.
3. Signal peptides.
4. Transmembrane domains.
5. Post-translational modifications.
6. Motifs & domains.
7. Secondary structure.
8. Other resources.

ExPASy - http://www.expasy.ch/

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) protein and proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures. Besides the tools that we will introduce in this manual there are many other applications available at this website that you should take some time to have a look at.

1) Physico-chemical properties:

ProtParam tool – http://www.expasy.ch/tools/protparam.html

Calculates lots of physico-chemical parameters of a protein sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY)

Example: Human BRCA 1

You can paste the gene sequence from the course Website.

- At ExPASy → "Proteomics and sequence analysis tools" → "Primary structure analysis".
- Click on the "ProtParam" link.
- Paste your sequence in the box provided
- The sequence must be written using the one letter amino acid code:
- Press the "Compute parameters" button.
- The output for this sequence is shown below.

Number of amino acids: 1863
Molecular weight: 207720.8
Theoretical pI: 5.29
Amino acid composition:
Ala (A) 84 4.5%

Arg (R) 76 4.1%
Etc etc
Thr (T) 111 6.0%
Trp (W) 10 0.5%
Tyr (Y) 31 1.7%
Val (V) 101 5.4%
Asx (B) 0 0.0%
Glx (Z) 0 0.0%
Xaa (X) 0 0.0%

Total number of negatively charged residues (Asp + Glu): 283
Total number of positively charged residues (Arg + Lys): 213
Atomic composition:
Carbon C 8908
Hydrogen H 14246
Nitrogen N 2554
Oxygen O 3014
Sulfur S 74
Formula: $C_{8908}H_{14246}N_{2554}O_{3014}S_{74}$
Total number of atoms: 28796
Extinction coefficients:
Conditions: 6.0 M guanidium hydrochloride
0.02 M phosphate buffer
pH 6.5
Extinction coefficients are in units of M^{-1} cm^{-1}.
The first table lists values computed assuming ALL Cys residues appear as half cystines, whereas the second table assumes that NONE do.

```
276 278 279 280 282
nm nm nm nm nm
Ext. coefficient 102140 102194 100935 99220 95840
Abs 0.1% (=1 g/l) 0.492 0.492 0.486 0.478 0.461
276 278 279 280 282
nm nm nm nm nm
Ext. coefficient 98950 99400 98295 96580 93200
Abs 0.1% (=1 g/l) 0.476 0.479 0.473 0.465 0.449
```

Estimated half-life:
The N-terminal of the sequence considered is M (Met).
The estimated half-life is: 30 hours (mammalian reticulocytes, in vitro).
>20 hours (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).
Instability index:
The instability index (II) is computed to be 54.68
This classifies the protein as unstable.
Aliphatic index: 69.01
Grand average of hydropathicity (GRAVY): -0.785

2) Cellular localization:

PSORT - http://psort.nibb.ac.jp/form2.html

PSORT, a program to predict the subcellular localization sites of proteins from their amino acid sequences. This program makes use of the fact that proteins destined for particular subcellular localizations have distinct amino acid properties particularly in their N-terminal regions. These properties can be used to predict whether a protein is localized in the cytoplasm, nucleus, mitochondria, or is retained in the ER, or destined for the lysosome (vacuolar) or the peroxisome. There is a detailed page of output that we can probably ignore. At the end of the output the percentage likelihood of the subcellular localization is given. If you want to learn more about the output and how subcellular localization is determined please see the user manual at:

http://psort.nibb.ac.jp/helpwww2.html

Example: Human ETS-1 protein.

- At ExPASy → "Post-translational modification prediction".
- Click on the "PSORT" link.
- For animal/yeast sequences click the link to "PSORT II Prediction".
- Paste your sequence in the box provided.
- The sequence must be written using the one letter amino acid code:
- Press the submit button.
- The output for this sequence is shown below.

There are a number parameters measured by this program which you can read about as links from the output file. By scrolling to the bottom of the output you can see the probability that this sequence is nuclear, cytoplasmic, peroxisomal, vacuolar or cytoskeletal. PSORT predicts that ETS-1 is nuclear with a high probability. The fact that ETS-1 is localized in the nucleus has been previously experimentally determined.

Results of Subprograms

```
PSG: a new signal peptide prediction method
N-region: length 8; pos.chg 2; neg.chg 1
H-region: length 6; peak value 1.89
PSG score: -2.51
```

Results of the k-NN Prediction

```
k = 9/23
73.9 %: nuclear
13.0 %: cytoplasmic
4.3 %: peroxisomal
4.3 %: vacuolar
4.3 %: cytoskeletal
>> prediction for QUERY is nuc (k=23)
```

3) Signal peptides:

Proteins destined for secretion, operation with the endoplasmic reticulum, lysosomes and many transmembrane proteins are synthesized with leading (N-terminal) 13 – 36 residue signal peptides.

SignalP - http://www.cbs.dtu.dk/services/SignalP/

The SignalP WWW server can be used to predict the presence and location of signal peptide cleavage sites in your proteins. It can be useful to know whether your protein has a signal peptide as it indicates that it may be secreted from the cell. Furthermore, proteins in their active form will have their signal peptides removed, if you can determine the length of the signal peptide then you can calculate the size of the protein minus the signal peptide.

Example: Human Beta-defensin; sp|Q09753|BD01_HUMAN

At ExPASy → "Post-translational modification prediction".

Click on the "SignalP" link.

Paste your sequence in the box provided

The sequence must be written using the one letter amino acid code:

It is recommend that the N-terminal part only (not more than 50-70 amino acids) of the sequences is submitted. A longer sequence will increase the risk of false positives and make the graphical output difficult to read. The new version now automatically truncates input sequences.

Choose one or more group of organisms for the prediction by clicking the check-box next to the group(s):

If no groups are indicated, predictions from all three groups will be returned.

A graphical output (in Postscript format) of the prediction will be available, if the "Include graphics"-button is checked.

Press the "Submit sequence" button.

A WWW page will return the results when the prediction is ready. Response time depends on system load. The output for this sequence is shown below

C score = raw cleavage site score

The output score from networks trained to recognize cleavage sites vs. other sequence positions. Trained to be: High at position +1 after the cleavage site and low at all other positions.

S score = signal peptide score

The output score from networks trained to recognize signal peptide vs. non-signal-peptide positions. Trained to be: High at position before the cleavage site and low at all other positions.

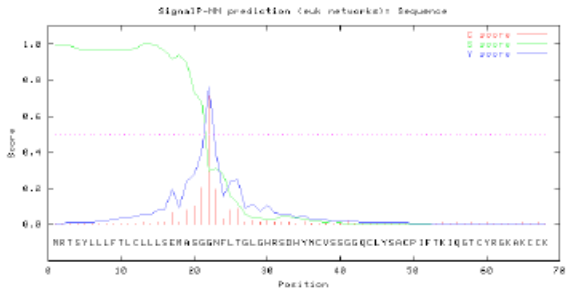
Y score = combined cleavage site score

The prediction of cleavage site location is optimized by observing where the C-score is high and the S-score changes from a high to a low value.

For each sequence, SignalP will report the maximal C, S, and Y scores, and the mean S-score between the N-terminal and the predicted cleavage site. These values are used to distinguish between signal peptides and non-signal peptides. If your sequence is predicted to have a signal peptide, the cleavage site is predicted to be immediately before the position with the maximal Y-score.

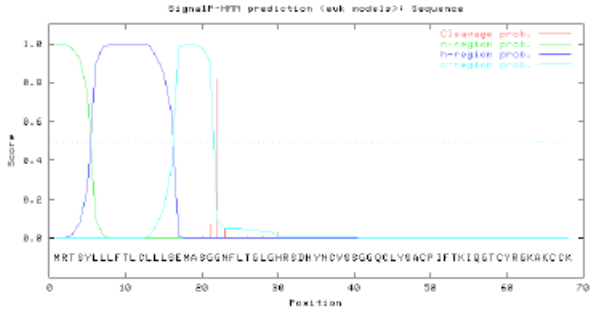
The Human beta-defensin protein has a predicted signal peptide from position 1 to 21 and a potential cleavage site exists between positions 21 and 22. These predictions correspond exactly to the SWISS-PROT annotation for this protein (accession Q09753).

SignalP-NN result:




```
# data
>Sequence length = 68
# Measure Position Value Cutoff signal peptide?
max. C 22 0.710 0.32 YES
max. Y 22 0.761 0.33 YES
max. S 14 0.998 0.87 YES
mean S 1-21 0.943 0.48 YES
D 1-21 0.852 0.43 YES
# Most likely cleavage site between pos. 21 and 22: ASG-GN
```

SignalP-HMM result:



```
# data
>Sequence
Prediction: Signal peptide
Signal peptide probability: 1.000
Signal anchor probability: 0.000
Max cleavage site probability: 0.818 between pos. 21 and 22
```

4) Transmembrane domains:

Tmpred - http://www.ch.embnet.org/software/TMPRED_form.html

The TMpred program makes a prediction of membrane-spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase, a database of naturally occurring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring. The presence of transmembrane domains is an indication that the protein is located on the cell surface.

Example: Human chemokine receptor 4 protein sequence NP_003458.1

At ExPASy → "Topology prediction".

Click on the link to Tmpred.

Paste your sequence in the box provided in one of the supported formats e.g.

plain text, SwissProt_ID or AC, etc.

You may change the minimal and maximal length of the hydrophic part of the transmembrane helix but unless you have reason to do so you should accept the defaults i.e. 17 and 33. ~22 residues is the same length as the width of a lipid bilayer.

Click the "Run Tmpred" button to start the search.

The output is given in 3 parts 1, 2 and 3 (see below).

Part 1: lists all the significant predictions of possible transmembrane helices

in this case there are 7 helices predicted but at this stage we do not know the orientation of the helices so there are 2 tables, the first with the helices orientated from the inside to the outside and vice versa for the second.

Part 2: shows which inside->outside helices correspond to the outside -> inside helices and indicates which orientation is most likely.

Part 3: proposes the strongly preferred model for the transmembrane domain structure of the protein and also an alternative model.

A graphic of the prediction is also available (not shown here)

These predictions correspond well but not exactly to the SWISS-PROT annotation for this protein (accession P30991).

Tmpred output

Sequence: MEG...HSS, length: 352

Prediction parameters: TM-helix length between 17 and 33

---

1. Possible transmembrane helices

The sequence positions in brackets denominate the core region. Only scores above 500 are considered significant.

```
Inside to outside helices : 7 found
from to score center
39 ( 46) 62 ( 62) 1962 54
78 ( 85) 105 ( 103) 1623 95
114 ( 114) 133 ( 130) 1352 122
155 ( 157) 175 ( 173) 1716 165
204 ( 206) 223 ( 223) 2052 214
240 ( 240) 261 ( 259) 2840 251
286 ( 286) 305 ( 305) 1241 295
Outside to inside helices : 7 found
from to score center
47 ( 47) 63 ( 63) 2568 55
78 ( 78) 96 ( 96) 1331 86
111 ( 114) 132 ( 132) 1740 122
155 ( 157) 173 ( 173) 1197 165
204 ( 204) 223 ( 223) 2404 214
240 ( 242) 259 ( 259) 2037 251
283 ( 286) 305 ( 305) 1703 294
```

---

2. Table of correspondences

Here is shown, which of the inside->outside helices correspond to which of the outside->inside helices.
Helices shown in brackets are considered insignificant. A “+”-symbol indicates a preference of this orientation. A “++”-symbol indicates a strong preference of this orientation.

```
Inside->outside | outside->inside
39- 62 (24) 1962 | 47- 63 (17) 2568 ++
78- 105 (28) 1623 ++ | 78- 96 (19) 1331
114- 133 (20) 1352 | 111- 132 (22) 1740 ++
155- 175 (21) 1716 ++ | 155- 173 (19) 1197
```

```
204- 223 (20) 2052 | 204- 223 (20) 2404 ++
240- 261 (22) 2840 ++ | 240- 259 (20) 2037
286- 305 (20) 1241 | 283- 305 (23) 1703 ++
```

3. Suggested models for transmembrane topology

These suggestions are purely speculative and should be used with extreme caution since they are based on the assumption that all transmembrane helices have been found. In most cases, the Correspondence Table shown above or the prediction plot that is also created should be used for the topology assignment of unknown proteins.

2 possible models considered, only significant TM-segments used

```
--->  STRONGLY preferred model: N-terminus outside
7 strong transmembrane helices, total score : 14594
# from to length score orientation
1 47 63 (17) 2568 o-I
2 78 105 (28) 1623 I-o
3 111 132 (22) 1740 o-I
4 155 175 (21) 1716 I-o
5 204 223 (20) 2404 o-I
6 240 261 (22) 2840 I-o
7 283 305 (23) 1703 o-I
---->  alternative model
7 strong transmembrane helices, total score : 11172
# from to length score orientation
1 39 62 (24) 1962 I-o
2 78 96 (19) 1331 o-I
3 114 133 (20) 1352 I-o
4 155 173 (19) 1197 o-I
5 204 223 (20) 2052 I-o
6 240 259 (20) 2037 o-I
7 286 305 (20) 1241 I-o
```

5) Post-translational modifications:

After translation has occurred proteins may undergo a number of posttranslational modifications. These can include the cleavage of the pro- region to release the active protein, the removal of the signal peptide and numerous covalent modifications such as, acetylations, glycosylations, hydroxylations, methylations and phosphorylations. Posttranslational modifications such as these may alter the molecular weight of your protein and thus its position on a gel. There are many programs available for predicting the presence of posttranslational modifications, we will take a look at one for the prediction of type O-glycosylation sites in mammalian proteins. Remember these programs work by looking for consensus sites and just because a site is found does not mean that a modification definitely occurs.

NetOGlyc - http://www.cbs.dtu.dk/services/NetOGlyc-2.0/

Prediction of type O-glycosylation sites in mammalian proteins. This program works by comparing the input sequence to a database of known and verified mucin type O-glycosylation sites extracted from O-GLYCBASE.

Example: Human CD1D sp|P15813|CD1D_HUMAN

- At ExPASy → "Post-translational modification".

- Click on the link to "NetOGlyc".
- Paste your sequence in the box provided in FASTA format.
- Check "generate graphics" and click the submit button.
- The output for this program is shown below (graphics not shown).
- This program predicts potential O-glycosylation sites at Threonine 64 and Serine 214.

NetOGlyc 2.0 Prediction Results

```
Name: Sequence Length: 335
MGCLLFLLLWALLQAWGSAEVPQRLFPLRCLQISSFANSSWTRTDGLAWLGELQTHSWSNDSDTVRSLKPWSQGTFSDQQ 80
WETLQHIFRVYRSSFTRDVKEFAKMLRLSYPLELQVSAGCEVHPGNASNNFFHVAFQGKDILSFQGTSWEPTQEAPLWVN 160
LAIQVLNQDKWTRETVQWLLNGTCPQFVSGLLESGKSELKKQVKPKAWLSRGPSPGPGRLLLVCHVSGFYPKPVWVKWMR 240
GEQEQQGTQPGDILPNADETWYLRATLDVVAGEAAGLSCRVKHSSLEGQDIVLYWGGSYTSMGLIALAVLACLLFLLIVG 320
FTSRFKRQTSYQGVL
.....................................................................T..... 80
.....................................................................S..... 160
.....................................................................S..... 240
.....................................................................S..... 320
...............
```

Name	Residue	No.	Potential	Threshold	Assignment
Sequence	Thr	42	0.0611	0.6493	.
Sequence	Thr	44	0.0087	0.6573	.
Sequence	Thr	55	0.0117	0.6491	.
Etc	etc				
Sequence	Thr	248	0.0131	0.5840	.
Sequence	Thr	260	0.0089	0.6578	.
Sequence	Thr	266	0.0224	0.6957	.
Sequence	Thr	300	0.0147	0.7357	.
Sequence	Thr	322	0.0480	0.7096	.
Sequence	Thr	329	0.0639	0.6021	.

Name	Residue	No.	Potential	Threshold	Assignment
Sequence	Ser	18	0.0161	0.6211	.
Sequence	Ser	34	0.0044	0.6673	.
Sequence	Ser	35	0.0048	0.6717	.
Sequence	Ser	39	0.0337	0.6118	.
Sequence	Ser	40	0.0013	0.6521	.
Sequence	Ser	57	0.0065	0.6484	.
Etc	etc				
Sequence	Ser	284	0.0005	0.6401	.
Sequence	Ser	285	0.0082	0.6389	.
Sequence	Ser	298	0.0003	0.6778	.
Sequence	Ser	301	0.0007	0.6924	.
Sequence	Ser	323	0.0003	0.6441	.
Sequence	Ser	330	0.0052	0.6277	.

Note: The new version of this server does not predict these sites. This is a good lesson in the evolving nature of these servers and why validation at more than one is a good idea.

6) Motifs and Domains

If you want to determine the function of a protein the first tool of choice is homology searching (see day 4). Unless this finds you a match with a well characterized protein comprehending the entire length of yours you should look for motifs and domains in your protein. To determine if your protein sequence contains known motifs or conserved domain structures you should search the protein against one of the motif or profile databases. There are many of these available but we will discuss ProfileScan

(now called myHits), which allows you to search both the Prosite and Pfam databases simultaneously. See the documentation for more details.

ProfileScan - http://hits.isb-sib.ch/cgi-bin/PFSCAN

Example: Human CFTR sp|P13569|CFTR_HUMAN

- Go to the URL above
- Paste your sequence in the box provided.
- The sequence must be written using the one letter amino acid code:
- Tick the motif databases you wish to search, other parameters should be OK.
- Press the "scan" button.

The output for this program is too large to show here, but it gives lots of detail about motifs in the CFTR protein identifying potential: ABC transporters family signature; ATP/GTP-binding site motif A (P-loop); Protein kinase C phosphorylation sites; N-glycosylation sites; Casein kinase II phosphorylation site; N-myristoylation sites; cAMP- and cGMP-dependent protein kinase phosphorylation site; Bipartite nuclear localization signal; NACHT-NTPase domain profile; Guanylate kinase domain profile etc.

Remember that these programs only tell you are that there is a motif present and thus there is the potential for these modifications and functions to occur. It is up to you to determine experimentally which are real but at least you now know what to look for.

7) Secondary Structure Prediction

If protein structure, even secondary structure, can be accurately predicted from the now abundantly available gene and protein sequences, such sequences become immensely more valuable for the understanding of drug-design, the genetic basis of disease, the role of protein structure in its enzymatic, structural, and signal transduction functions, and basic physiology from molecular to cellular, to fully systemic levels. In short, the solution of the protein structure prediction problem (and the related protein folding problem) will bring on the second phase of the molecular biology revolution (Munson et al., 1994).

JPRED - http://www.compbio.dundee.ac.uk/~www-jpred/submit.html

Jpred is an Internet web server that takes either a protein sequence or a multiple alignment of protein sequences, and predicts secondary structure. It works by combining a number of modern, high quality prediction methods to form a consensus. Please be aware that secondary structure prediction is an extremely complex problem that is under intensive research and we are still at a relatively primitive stage. We cannot discuss the details of protein secondary structure here but if you are interested in this area we recommend that you take a look at any major biochemistry textbook. Essentially protein secondary structure consists of 3 major conformations; the α Helix, the β pleated sheet and the coil conformation.

Example: Human alpha 1 hemoglobin. NP_000549.1

- At the ExPASy → "Secondary structure prediction".
- Click on the link to JPRED.
- Click "Prediction".
- Paste your sequence in the box provided.
- The defaults are OK.
- Click "Run secondary structure predictions!"
- Point 4 on the submission page allows you to deselect the BLAST search against PDB (Protein Data Bank). If your sequence already has had its structure predicted or experimentally determined it will be in here and you can follow the link to PDB for information on the structure of your protein.
- If your protein is in PDB you can view your protein secondary structure using RasMol (To download RasMol see the course website for a link).
- Once you have RasMol running you can open your structure in it a view it using a number of different options.

Otherwise continue with prediction

- The program may take a long time so you can save a bookmark and return to your results later or choose to have your results e-mailed to you.
- There are a number of options to view the output, view your output in HTML format (option 4).
- The complete output is too large to show here (see webpage).
- Scroll down through the output until you get to "Jpred" output. The line of output beside this is the consensus secondary structure for your sequence. H= Helices E= strands C= coils.

A Few Other Useful Tools at ExPASy

FindMod

Predicts potential protein post-translational modifications (PTM) and find potential single amino acid substitutions in peptides. The experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified SWISS-PROT/TrEMBL entry or from a user-entered sequence, and mass differences are used to better characterise the protein of interest.

NetPhos:

The NetPhos WWW server produces neural network predictions for serine, threonine and tyrosine phosphorylation sites in eukaryotic proteins.

Sulfinator:

Predicts tyrosine sulfation sites in protein sequences. Tyrosine sulfation is an important post-translational modification of proteins that go through the secretory pathway.

REP:

Searches a protein sequence for a collection of repeats such as leucine rich repeats and many others.

Other Resources for Protein Sequence Analysis

1) Protein Prospector at UCSF - http://prospector.ucsf.edu/

MS-Digest: A protein digestion tool from the UCSF Mass Spectrometry Facility that performs an in silico enzymatic digestion of a protein sequence, and calculates the mass of each peptide.

MS-Product: A tool from the UCSF Mass Spectrometry Facility that calculates the possible fragment ions resulting from fragmentation of a peptide in a mass spectrometer. Fragmentation possibilities for post-source decay (PSD), high-energy collision-induced dissociation (CID), and low-energy CID processes may be calculated.

2) Pasteur Institute - http://bioweb.pasteur.fr/seqanal/protein/intro-uk.html

Antigenic: finds antigenic sites in proteins.

Helixturnhelix: reports nucleic acid binding motifs in your protein of interest.

Accessing Completed Genomes

TOPICS:

1. **GeneDB.**
2. **TigrDB.**
3. **Ensembl.**
4. **NCBI Genomic Biology.**

Accessing Genomic Sequences:

There is no one resource available on the web that allows you to access all the available genomes. In this course we will take a look at 3 sites for accessing most of the genomic information that is available out there. These sites often contain similar information and it may be possible to get most of the information you require from just one of these sites, however, to get the maximum amount of information it is often worth having a look at all 3 of these sites. In this course we will primarily concentrate on accessing the Trypanosome, human bovine genome data, however, any of the examples that we describe can easily be applied to any of the available species. Remember that most of the genomes are still in a draft state and are subject to change as more sequence becomes available.

GeneDB - http://www.genedb.org/

What is GeneDB?

Funded as part of the Wellcome Trust Functional Genomics Development Initiative, the GeneDB project is aiming to develop and maintain curated database resources for three organisms: *Schizosaccharomyces pombe*, which has been completely sequenced, and the kinetoplastid protozoa *Leishmania major* and *Trypanosoma brucei*, whose genome sequences have yet to be completed. The goals are to capture, store and manage data for integration with emerging functional genomics and proteomics projects and to provide an easy-to-use, user-friendly interface, including a variety of graphical displays. It is envisaged that the generic database structure will subsequently be adopted to integrate datasets for other organisms, both prokaryotic and eukaryotic, that have been sequenced by the Sanger Institute Pathogen Sequencing Unit.

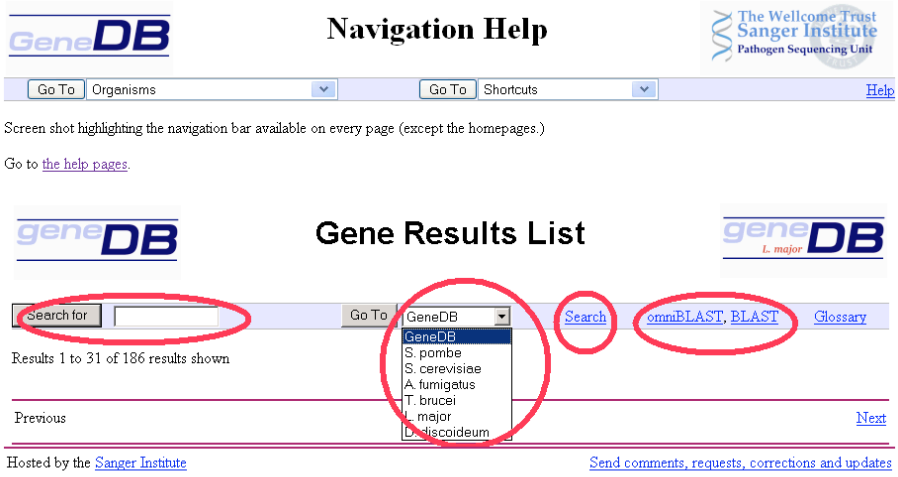
To this extend, datasets for *Saccharomyces cerevisiae* as well as the filamentous fungus *Aspergillus fumigatus* are already available through GeneDB. The database has been developed through close collaboration between Sanger Institute software developers, on-site organism specific curators and representatives of the research communities. The data within geneDB are manually annotated and curated, frequently updated and, because of the structured annotation and use of controlled vocabulary, easy to precisely query. The database is under constant review and new functionality will be added as it evolves.

What are the various ways to search GeneDB?

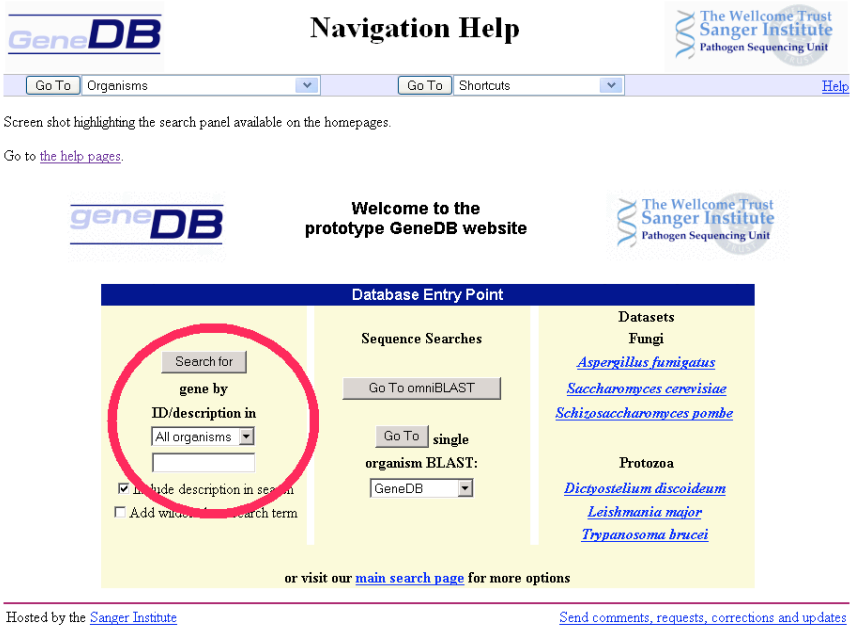
GeneDB provides users with the following information, functionality and research tools. The following are descriptions of ways to search GeneDB, where links will take you to the relevant areas of the database or to example pages. All the relevant search pages are available from a database entry point on both the GeneDB homepage and the individual organism homepages.

Menu Bar

To aid in the navigation of the site a menu bar is available on the pages within GeneDB. Most of the options available on GeneDB frontpages are featured together with a comprehensive glossary of useful terms and databases. The menu bar has a gene search box, a drop down menu for the other organisms within GeneDB, Blast search and a link to the main search page. Also available are links to the GeneDB and organism frontpages via the prominant GeneDB logos above the menu bar.

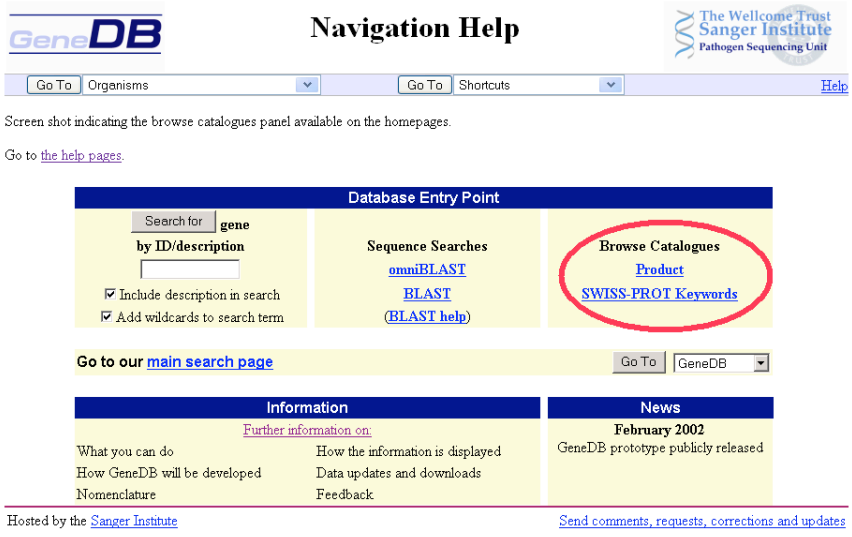


Searching for a gene by name or synonym



This option is also available on the menu bar of each gene page. Entering a gene name or synonym will lead either directly to the relevant gene page (eg dld1 in *S. pombe*) if a specific unique term is used or to a list of genes including that term (eg *kinase* in *T. brucei* and *L. major*) if a wild card is used. A list of genes will provide links to each relevant gene page.

Browsing by catalogues



The list of browsable terms includes

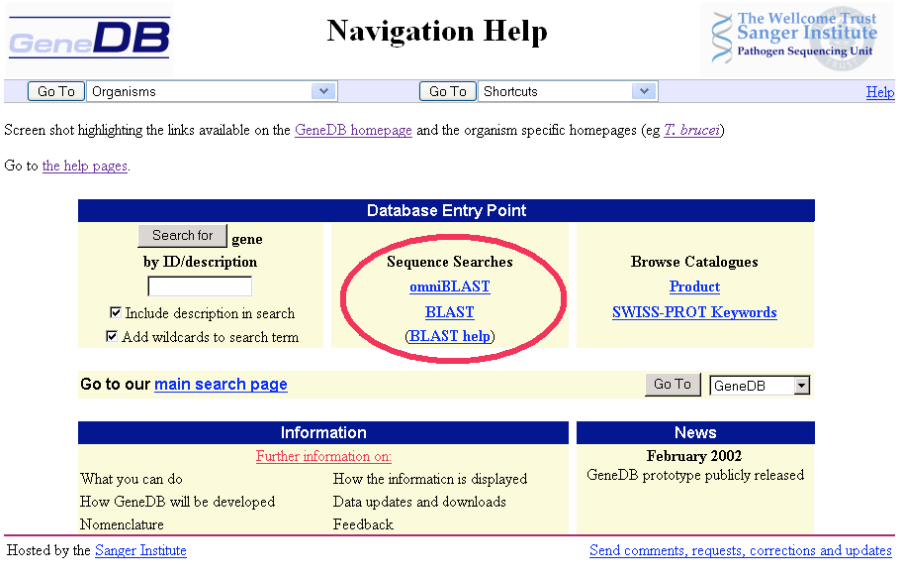
products/description: This brings up a gene product list with links to relevant gene pages for each product

SWISS-PROT keywords: This is a browsable list of SWISS-PROT keywords assigned by SWISS-PROT curators to a given protein linking to the relevant gene or gene list pages

GO (Gene Ontology) list: This allows the user to search for genes by classification of their respective products into molecular function, biological process and cellular component using the controlled vocabulary defined by the GO consortium.

Pfam (domain data): This is a list of protein domain families defined by sequence alignments and hidden Markov models.

Sequence searching using BLAST



The BLAST and omniBLAST links lead to self-explanatory search pages, allowing users to paste in any nucleotide or amino acid sequence and compare it for similarity to any sequence within the database. Results are returned either directly or by e-mail.

Full text search

GeneDB
Navigation Help
The Wellcome Trust Sanger Institute Pathogen Sequencing Unit
Go To Organisms
Go To Shortcuts
Help
Screen shot highlighting the full-text search available on the compact search page and the longer, more friendly search page
Go to the help pages.
GeneDB
GeneDB Search Page
The Wellcome Trust Sanger Institute Pathogen Sequencing Unit
Go To GeneDB homepage
Search
omniBLAST
Glossary
This is a compact query form for experienced users. Our longer search page has more explanations of the options available.
If you want to do sequence similarity searches, please go to our main omniBLAST page.
Quick Search
Search for gene name, systematic id
Include description in search
Search
Full text search (sitewide)
Go
Quick Browse By...
Products
SWISS-PROT Keywords
GO Component Assignments
GO Process Assignments
GO Functional Assignments
Not yet available
Pfam Assignments
S. pombe
Browse
Hosted by the Sanger Institute
Send comments, requests, corrections and updates

Search the whole site for any text reference using a web-based search engine. This option is available via the main search page link in the database entry point of each front page. A full text search box allows for searches across all the organisms within the database or just for a selected organism chosen from a drop down menu.

What kinds of information are in GeneDB?

Central to GeneDB are the gene pages, providing a comprehensive annotation of genes within each organism with:

3. Access DNA and protein sequences of protein coding genes with the ability of sending sequences straight to the associated BLAST servers
4. Predicted peptide properties (including signal peptide and transmembrane predictions)
5. Similarity information (EMBL, SWISS-PROT including annotation)
6. Gene ontology (GO) annotation
7. Summary of up-to-date protein domain and motif searches (InterPro, Pfam, PRINTS, PROSITE, BLOCKS, SMART)
8. Literature links
9. SWISS-PROT annotation

TigrDB - http://www.tigr.org/db.shtml

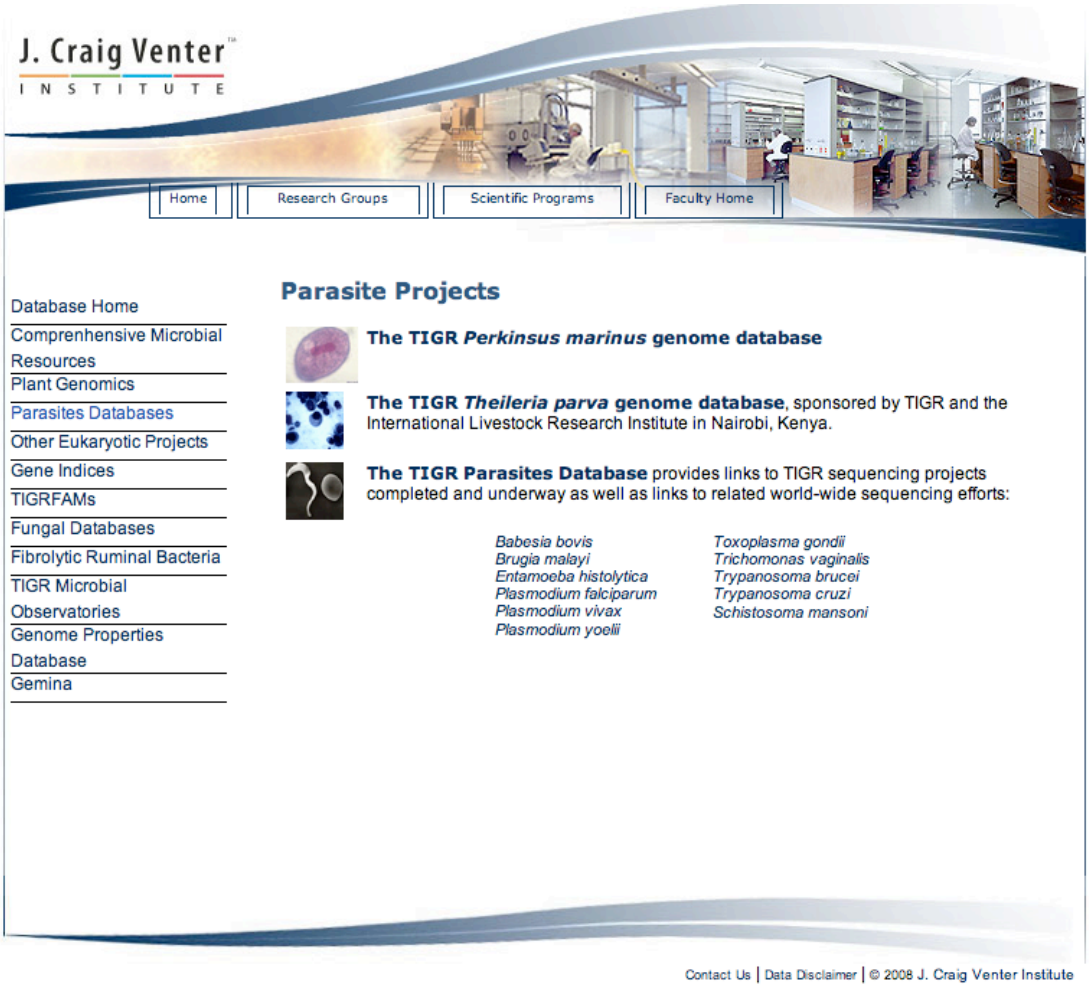
TIGR's Genome Projects are a collection of curated databases containing DNA and protein sequence, gene expression, cellular role, protein family, and taxonomic data

for microbes, plants and humans. Anonymous FTP access to sequence data is also provided. Please read the disclaimer regarding use of data. The TIGR clone distribution policy is available for viewing. There are several databases available on the TIGR database website: http://www.tigr.org/db.shtml. Several types of databases are found on the TIGR website. One is for completed and unfinished parasite genome sequences, sequenced at TIGR. A second valuable type of databases are the TIGR Gene Index.

J. Craig Venter
INSTITUTE
Home
Research Groups
Scientific Programs
Faculty Home
Database Home
Comprenhensive Microbial Resources
Plant Genomics
Parasites Databases
Other Eukaryotic Projects
Gene Indices
TIGRFAMs
Fungal Databases
Fibrolytic Ruminal Bacteria
TIGR Microbial Observatories
Genome Properties Database
Gemina
Databases
TIGR's Genome Projects are a collection of curated databases containing DNA and protein sequence, gene expression, cellular role, protein family, and taxonomic data for microbes, plants and humans. The access to the data is facilitated by TIGR's Internet2 high-speed research network connection which is supported in part by the National Science Foundation under grant ANI-0333537. Anonymous FTP access to sequence data is also provided. Please read the disclaimer regarding use of data. The TIGR clone distribution policy is available for viewing.
Contact Us | Data Disclaimer | © 2008 J. Craig Venter Institute

Parasites Databases

Under the Parasites Databases, several completed parasite databases are found including data on several uncompleted genomes i.e. *Trypanosoma brucei*.



Data in the database can be accessed through several methods:

Gene Name Search Text search of the putative identifications in the Gene Identification Table.

Locus Search Obtain a report on a predicted coding region by locus number.

Sequence Search Provides searching of nucleotide or peptide sequences against predicted coding regions or the chromosomes.

HMM Search Search a sequence against protein family based HMMs

View Chromosomes Browse the chromosomes or retrieve a table of clones sorted by chromosome.

Gene Index project

The promise of genome projects has been a complete catalogue of genes in a wide range of organisms. While genome projects have been successful in providing reference genome sequences, the problem of finding genes and their variants in genomic sequence remains an ongoing challenge.

The sequencing of Expressed Sequence Transcripts (ESTs), fragments of genes that have been copied from DNA to RNA, provides the most comprehensive evidence for the existence of genes and their structure.

The goal of The Gene Index Project is to use the available EST and gene sequences, along with the reference genomes wherever available, to provide an inventory of likely genes and their variants and to annotate these with information regarding the functional roles played by these genes and their products. In addition, they are attempting to use these catalogues to find links between genes and pathways in different species and to provide lists of features within completed genomes that can aid in the understanding of how gene expression is regulated.

DFCI T.brucei Gene Index

The DFCI T. brucei Gene Index integrates research data from international T. brucei EST sequencing and gene research projects. The ultimate goal of the DFCI Gene Index projects, including TbGI, is to represent a non-redundant view of all T. brucei genes and data on their expression patterns, cellular roles, functions, and evolutionary relationships. Data can be accessed through several means:

BLAST search TC sequences based on sequence similarity

Identifiers or Keywords search TC reports using TC identifiers, GB accessions or keywords

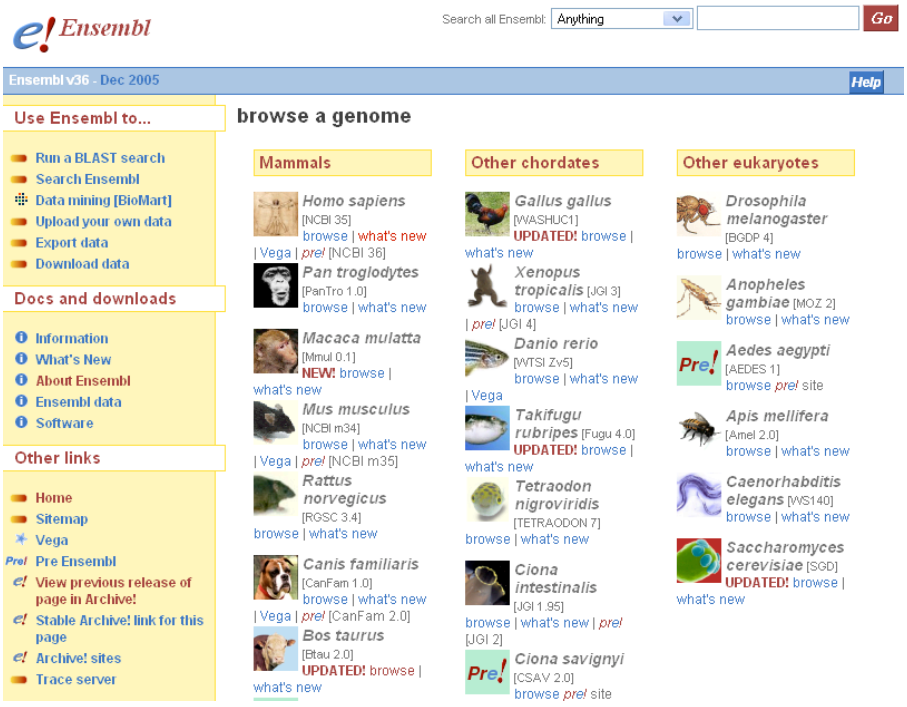
TC Annotator list all TC annotation

EST Annotator list all EST annotation

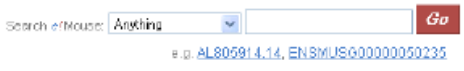
And several other.

ENSEMBL - http://www.ensembl.org/

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on eukaryotic genomes. A wide range of genomes are available.


Ensembl
Search all Ensembl: Anything Go
Ensembl v36 - Dec 2005
Help
Use Ensembl to...
Run a BLAST search
Search Ensembl
Data mining [BioMart]
Upload your own data
Export data
Download data
Docs and downloads
Information
What's New
About Ensembl
Ensembl data
Software
Other links
Home
Sitemap
Vega
Pre Ensembl
View previous release of page in Archive!
Stable Archive! link for this page
Archive! sites
Trace server
browse a genome
Mammals
Homo sapiens [NCBI 35] browse | what's new | Vega | pre! [NCBI 36]
Pan troglodytes [PanTro 1.0] browse | what's new
Macaca mulatta [Mmul 0.1] NEW! browse | what's new
Mus musculus [NCBI m34] browse | what's new | Vega | pre! [NCBI m35]
Rattus norvegicus [RGSC 3.4] browse | what's new
Canis familiaris [CanFam 1.0] browse | what's new | Vega | pre! [CanFam 2.0]
Bos taurus [Btau 2.0] UPDATED! browse | what's new
Other chordates
Gallus gallus [WASHUC1] UPDATED! browse | what's new
Xenopus tropicalis [JGI 3] browse | what's new | pre! [JGI 4]
Danio rerio [WTSI Zv5] browse | what's new | Vega
Takifugu rubripes [Fugu 4.0] UPDATED! browse | what's new
Tetraodon nigroviridis [TETRAODON 7] browse | what's new
Ciona intestinalis [JGI 1.95] browse | what's new | pre! [JGI 2]
Ciona savignyi [CSAV 2.0] browse pre! site
Other eukaryotes
Drosophila melanogaster [BGDP 4] browse | what's new
Anopheles gambiae [MOZ 2] browse | what's new
Aedes aegypti [AEDES 1] browse pre! site
Apis mellifera [Amel 2.0] browse | what's new
Caenorhabditis elegans [WS140] browse | what's new
Saccharomyces cerevisiae [SGD] UPDATED! browse | what's new

- Click on one of the species to access the genomic information e.g. Mouse.
- To find your gene of interest you can enter in the empty box (top right-hand corner of page), the gene symbol, gene accession number, mRNA accession number, SwissProt accession number, EST accession number etc.



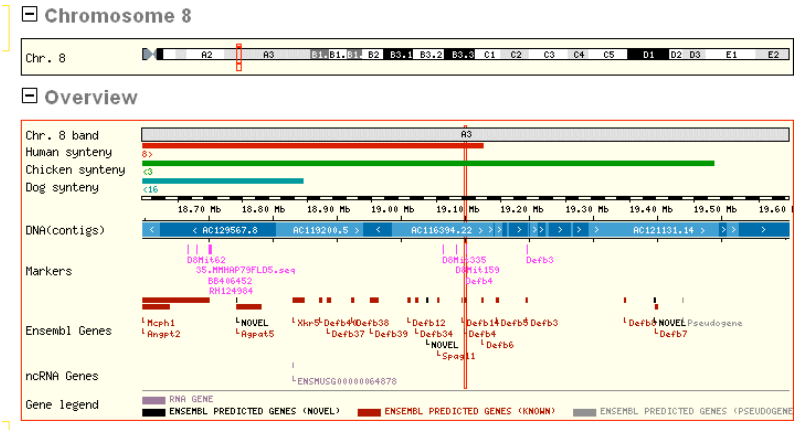
- You can also access the genome by chromosome number.
- There are a number of useful links located in the top left panel of the page

- Search Ensembl - BLAST/SSAHA- BLAST your sequence against the genome. SSAHA is similar to BLAT.
- Data mining [BioMart] – Allows the download of a large datasets e.g. all the genes on a chromosome, the entire genome etc.
- Upload your own data
- Export data - If you have an Ensembl I.D. for your gene you can download its sequence from here.
- Download data – ftp download of data (mostly used for large bioinformatics projects).

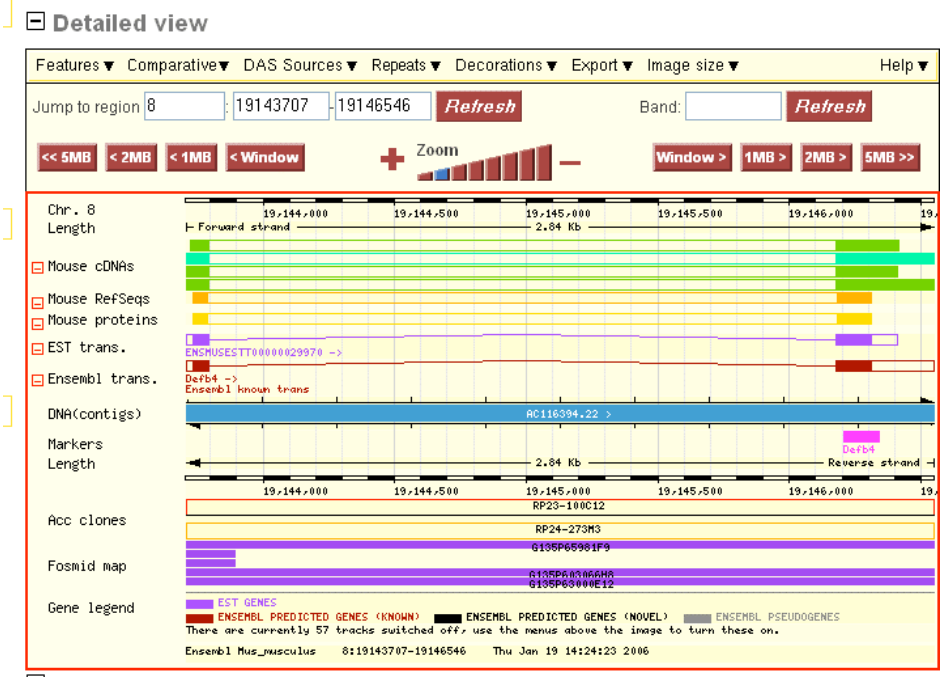
Example: Mouse beta-defensin 4 (defB4).

- Use the pull-down "Anything" menu to select "Gene" and type the gene RefSeq symbol in the empty box (defB4).
- Click "Lookup".
- This will take you to a query results page. In this case there is only one hit but sometimes you will have to look through a number of entries to find what you are looking for.
- Click on the **EnsEMBL Gene:** ENSMUSG00000059230 link for information on your gene such as its sequence, structure, domains that it contains etc.
- Click on the link to "Genomic Location" to display the gene in the genome similar to the UCSC browser.
- There are 4 major views displayed

- Chromosome – highlights position on the chromosome
- Overview – shows genes surrounding gene of interest on chromosome band
- Detailed View – more detailed view of your gene.
- Basepair View – displays sequence, translated sequence and restriction sites.



- On the detailed view you can use the pull down menus to "Features", "DAS Sources" & decorations to choose what details you wish to display.
- The "Export" menu can be used to download the genomic sequence or features of this area in the genome (E.g. list of genes).
- As with the UCSC browser you can zoom in or out of this region of genome or move along the genome using the "window buttons.



NCBI - http://www.ncbi.nlm.nih.gov/Genomes/index.html

Many of you will be familiar with the National Center for Biotechnology Information (NCBI) website which has many very useful resources including; Entrez, PubMed, Genbank, BLAST, OMIM. Today we will see how to use the NCBI site to interrogate the genomic sequences that are available there.

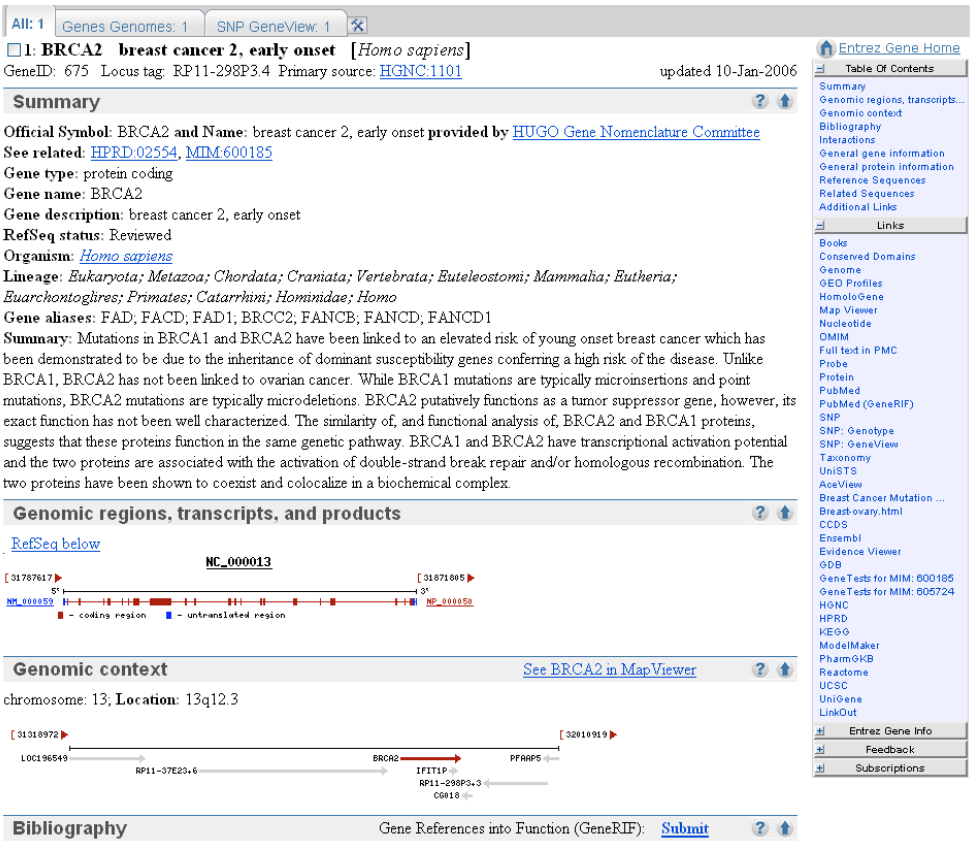
The NCBI site provides a good starting point for accessing the widest range of eukaryotic and microbial genomes. Many of these genomes will have their own dedicated sites located at other websites, but the NCBI site will provide links to them.

Accessing the Human Genome:

To access the human genome - go to the URL above and click on the link to "Human 'G' button". This page provides a number of links such as a link to BLAST where you can search your sequence against the human genome. You can also browse the genome by chromosome by clicking on one of the chromosomes. The best way to access the genome if you have a particular gene of interest is to search for your gene in Entrez Gene. Entrez Gene provides a single query interface to curated sequence and

descriptive information about genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers, UniGene clusters, homology, map locations, and related web sites.

- Follow the "Gene Database" link on the Human Genome Resources page.
- At the top of the page search Entrez Gene by entering your gene name (full name, abbreviation or accession number) in the box and "Go". (Example: BRCA2)
- This brings up a results page that matches the query for some reason.
- You can use the limits section to limit your search by various criteria such as organism.
- Click on BRCA2 (i.e. GeneID 675) to take you to the Entrez Gene page for that gene.


All: 1
Genes Genomes: 1
SNP GeneView: 1
1: BRCA2 breast cancer 2, early onset [Homo sapiens]
GeneID: 675 Locus tag: RP11-298P3.4 Primary source: HGNC:1101
updated 10-Jan-2006
Summary
Official Symbol: BRCA2 and Name: breast cancer 2, early onset provided by HUGO Gene Nomenclature Committee
See related: HPRD:02554, MIM:600185
Gene type: protein coding
Gene name: BRCA2
Gene description: breast cancer 2, early onset
RefSeq status: Reviewed
Organism: Homo sapiens
Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae; Homo
Gene aliases: FAD; FACD; FAD1; BRCC2; FANCB; FANCD; FANCD1
Summary: Mutations in BRCA1 and BRCA2 have been linked to an elevated risk of young onset breast cancer which has been demonstrated to be due to the inheritance of dominant susceptibility genes conferring a high risk of the disease. Unlike BRCA1, BRCA2 has not been linked to ovarian cancer. While BRCA1 mutations are typically microinsertions and point mutations, BRCA2 mutations are typically microdeletions. BRCA2 putatively functions as a tumor suppressor gene, however, its exact function has not been well characterized. The similarity of, and functional analysis of, BRCA2 and BRCA1 proteins, suggests that these proteins function in the same genetic pathway. BRCA1 and BRCA2 have transcriptional activation potential and the two proteins are associated with the activation of double-strand break repair and/or homologous recombination. The two proteins have been shown to coexist and colocalize in a biochemical complex.
Genomic regions, transcripts, and products
RefSeq below
NC_000013
Genomic context
See BRCA2 in MapViewer
chromosome: 13; Location: 13q12.3
BRCA2
Bibliography
Gene References into Function (GeneRIF): Submit
Entrez Gene Home
Table Of Contents
Summary
Genomic regions, transcripts...
Genomic context
Bibliography
Interactions
General gene information
General protein information
Reference Sequences
Related Sequences
Additional Links
Links
Books
Conserved Domains
Genome
GEO Profiles
HomoloGene
Map Viewer
Nucleotide
OMIM
Full text in PMC
Probe
Protein
PubMed
PubMed (GeneRIF)
SNP
SNP: Genotype
SNP: GeneView
Taxonomy
UniSTS
AceView
Breast Cancer Mutation ...
Breast-ovary.html
CCDS
Ensembl
Evidence Viewer
GDB
GeneTests for MIM: 600185
GeneTests for MIM: 605724
HGNC
HPRD
KEGG
ModelMaker
PharmGKB
Reactome
UCSC
UniGene
LinkOut
Entrez Gene Info
Feedback
Subscriptions

Starting at the top of the page:

- A graphic of the BRCA2 transcript is shown including the intron-exon structure. You can click on this graphic to obtain the sequence.
- This is followed by a graphic showing BRCA2 in its genomic context i.e. what genes are located around it.
- This is followed by various information on the gene including;

Gene aliases – other names for the gene.

Summary - written by staff of the NCBI RefSeq group describing the function, localization, and sequence properties of the gene and its products.

Bibliography – a detailed list of PubMed entries for the gene.

Interactions – What other genes/proteins are known to interact with BRCA2.

A General Gene Information Section – includes the official gene symbol and name, gene ontology details, homology with mouse and rat, etc.

There is also a link to the NCBI Map Viewer (see below).

NCBI Reference Sequences (RefSeq) - All RefSeq records created for a given locus are listed. Multiple records are distinguished by the brief description of the transcript variant. This section provides links to:

RefSeq nucleotide record (genomic and mRNA accessions have 'NG_' and 'NM_' prefixes, respectively).

RefSeq Product - protein record (the 'NP_' prefix).

Conserved domains found in the protein.

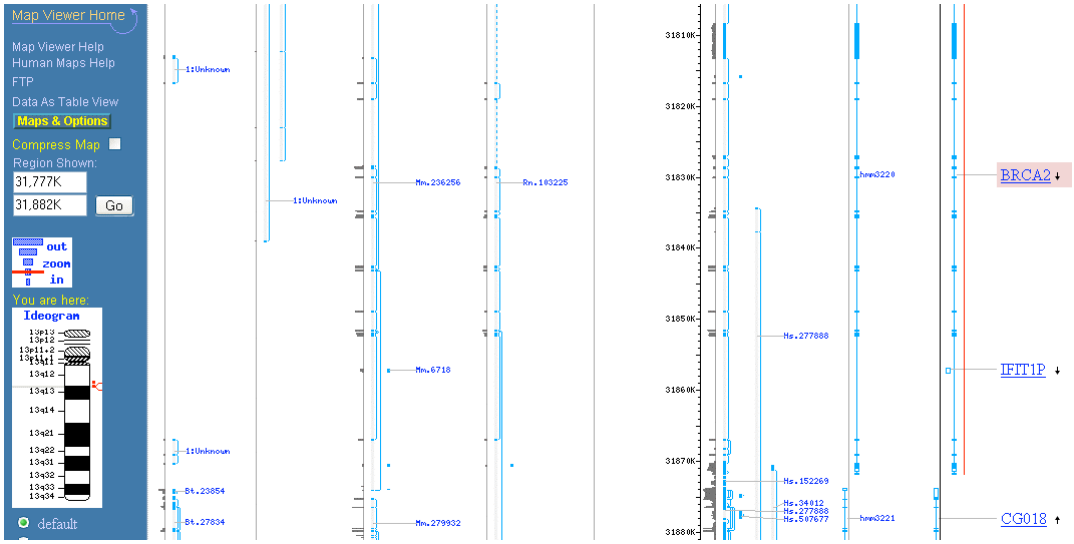
Related Sequences - A table of a subset of representative nucleotide and protein accessions for the locus. EST accession numbers are provided if no other sequence data are available to represent the locus.

Additional Links - This section names and provides links to additional sites that may contain information related to this locus such as OMIM, UniGene, etc.

MAP VIEWER:

This is the NCBI graphical display tool, which you can use to display the genomic context of your sequence. This tool is not as user friendly or as advanced as the UCSC or Ensembl browsers and we recommend that you use these to view the genome graphically where possible. Not all species are available at these sites so you may need to use Map viewer.

- Click on “Maps & Options” to choose which features you wish to display.
- Click on any of the genes, RNAs or Unigenes to get more information.
- You can download genomic sequence for the region selected using the “Download/View Sequence/Evidence” link.



Accessing the Other Genomes:
http://www.ncbi.nlm.nih.gov/Genomes/index.html

Plant Genomes Central

The plant genomic effort has one technical hurdle relative to other genomic efforts. The range of plant genome size is very large; extending from approximately the same size as small animals to more than five times as large as human. At NCBI resources for many plant species are available including;

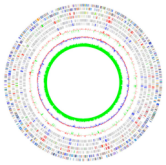
- *Arabidopsis thaliana* (thale cress)
- *Gossypium* (cotton)
- *Hordeum vulgare* (barley)
- *Lycopersicon esculentum* (tomato)
- *Medicago truncatula* (barrel medic)
- *Oryza sativa* (rice)
- *Solanum tuberosum* (potato)
- *Triticum aestivum* (bread wheat)
- *Zea mays* (corn)

Malaria Parasite

This resource provides data and information relevant to malaria genetics and genomics following the sequencing of the malaria parasite *Plasmodium falciparum* and one of its major vectors Anopheles gambiae genomes. These resources include;

- Organism specific sequence BLAST databases.

- Genome maps & linkage markers.
- Information about genetic studies.
- Links to other malaria web sites.
- Genetic data on related apicomplexan parasites.



Microbial Genomes

This resource provides links to the 279 (as of 07/11/2005) completely sequenced bacterial genomes (24 Archaea & 255 eubacteria). You can download information on the genome in a number of different formats

[T] - All proteins of the complete genome were searched against "nr" database. The detected homologs were classified into three taxonomic groups, Eukaryota, Eubacteria and Archaea in TaxTable.

[P] – Download the protein sequences from ProtTable.

[C] – Functional classifications are located in COG Table.

[D] - 3-D neighbors (proteins with sequence similarity to proteins with known 3D structure)

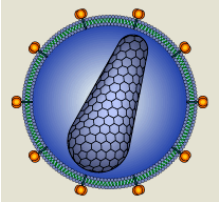
[L] – BLAST a sequence against the genome.

[S] - CDD search (list of conserved domains in proteins).

[F] – FTP data.

[R] – PubMed references.

For most of the genomes you can follow links to an organism-specific website with even further details (usually hosted by the sequencing consortium)



Retroviruses

Collection of resources at NCBI specifically designed to support the research of retroviruses. The resources include:

- Taxa-specific pages for HIV-1, HIV-2, SIV, HTLV, STLV.
- Genotyping tool - uses the BLAST algorithm to identify the genotype of a query sequence
- Alignment tool - global alignment of multiple sequences

- HIV-1 automatic sequence annotation - generates a report in GenBank format for one or more query sequences
- Genome maps - graphical representation of 50 retrovirus complete genomes

If you still can't find what you are looking for at any of these sites try:

The Institute for Genomic Research (TIGR) - http://www.tigr.org/

The Sanger Institute - http://www.sanger.ac.uk/

Some Other NCBI Resources:

Unigene - http://www.ncbi.nlm.nih.gov/UniGene/

UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location. The dataset is pretty comprehensive – for human there are:

54,576 sets total

In addition to sequences of well-characterized genes, hundreds of thousands novel expressed sequence tag (EST) sequences have been included. Consequently, the collection may be of use to the community as a resource for gene discovery. UniGene has also been used by experimentalists to select reagents for gene mapping projects and large-scale expression analysis.

It should also be noted that no attempt has been made to produce contigs or consensus sequences. There are several reasons why the sequences of a set may not actually form a single contig. For example, all of the splicing variants for a gene are put into the same set. Moreover, EST-containing sets often contain 5' and 3' reads from the same cDNA clone, but these sequences do not always overlap.

The NCBI genetic disease site - http://www.ncbi.nlm.nih.gov/disease/

This is rather a useful site, which classifies syndromes, diseases and conditions by sort: immune system, muscle and bone, signals, transporters, nervous system etc. You can browse through the hierarchy to find interesting diseases in your field of interest.

OMIM - http://www.ncbi.nlm.nih.gov/Omim/

The On-line Mendelian Inheritance in Man is a remarkable resource for all aspects of medical and clinical genetics. NCBI has a server that allows you to search this database.

Questions and Exercises

1) What contribution has Kirk Douglas made to medical/genetic research ?

2) What is the map-position of the gene involved in PKU ?

3) What happens when you search for Huntingdon ?

4) Better try Huntington ?

5) Any other genes where a key molecular biological flag is poly CAG repeats ?

6) For a female role model in science look up Julia Bell.

7) In what proportion of OMIM entries is "mental retardation" involved ?

Homology Searching

TOPICS

1. Introduction to homology searching.
2. BLAST.
3. FASTA.
4. Smith-Waterman.

Introduction

Perhaps the most widely used bioinformatics protocol is to search a database for sequences similar to a candidate sequence. Because of an implicit underlying hypothesis that if sequences are similar at some statistically significant level they share a common ancestor, this methodology is generally called homology searching. It is a useful tool because, if two sequences are similar, then they are likely to have a similar structure and if they have a similar structure they are likely to have a similar function. You can thus get important clues about the function of an as yet uncharacterized sequence.

There are several different algorithms for implementing a homology search, and each program will have a wide range of options and parameters to help you carry out a more informative type of search. The de facto standard for homology searching is the blast family of programs and this chapter will concentrate on them. You should note, however, that for searches with DNA sequences against DNA databases, the program Fasta is often more sensitive, if in general it will be a little slower. Smith-Waterman searches are generally more informative than either Blast or Fasta but very much slower.

Blast: http://www.ncbi.nlm.nih.gov/BLAST/

Blast is a finely tunable algorithm to search very large databases for homologues in a managable/finite time. It may be helpful to think that the complete human genome DNA comprises more than 3.2 * 10^9 bases. On a letter for letter basis this is the equivalent of about 8 complete Encyclopedia Britannicas. So the task of finding a sentence similar to the one you are now reading in such a forest of information is, shall we say, daunting. It is a 5 step process:

1. break the query sequence into a number of 'words' (typically 4 protein residues).
2. search the database for matches to these words.
3. the program builds on the "hits" by extending the alignment out on either side of the core word - these extended hits are called HSPs - high scoring segment pairs.

4. all the statistically significant segment pairs are sorted by some scoring criterion, so that the 'best' matches are presented first.
5. the significant matches are formally aligned to show where the homologous regions are.

Blast is not one program but a family of programs for carrying out different classes of search:

blastn: searches a DNA sequence against a DNA database such as EMBL, Genbank, or dbEST.

blastp: searches a protein sequence against a protein database such as Swissprot, or trembl (conceptual translations of the EMBL DNA database) or genpept (ditto for Genbank) or, most commonly, "nr" a non-redundant database which ideally contains one copy of every available sequence.

Then you have:

blastx: searches a DNA sequence (translated in all six reading frames) against a protein database.

tblastn: searches a protein sequence against a DNA database (translated in all six reading frames) – essential for searching EST databases.

and in the interests of completeness there is:

tblastx: searches a DNA sequence (translated in all six reading frames) against a DNA database (translated in all six reading frames).

See the Blast page at NCBI for details of other flavours of Blast programs.

Fasta:

The other widely used, although possibly not widely enough used, algorithm for doing homology searches against databases is Fasta, maintained by Bill Pearson in Virginia. You can carry out Fasta searches from: http://www.ebi.ac.uk/Tools/ this introductory course will not cover Fasta except to note that it is a) a little slower than blast b) it is the algorithm of choice if you have to search a DNA sequence against a DNA database.

Smith-Waterman:

These searches are very much more sensitive than either blast or fasta, but consequently take a much longer time to complete. Perhaps 20x slower than blast. One implementation of S-W is Blitz, which can be found on http://www.ebi.ac.uk/Tools/ the EBI homology server. In order to get S-W searches down to sensible times it is often carried out on Massively Parallel Computers.

Several Smith-Waterman algorithms are implemented on the BECA-ILRI High Performance computing cluster: http://hpc.ilri.cgiar.org/bwb/

Because for many biological searches blast will give you results that are a) good enough and b) returned in the shortest time, we will investigate that algorithm in more detail.

Options in blast.

Masking/filtering of less informative sequence motifs.

If your query sequence is protein you can "mask" regions of the protein that may give you confusing or biologically uninformative information. This masking can be of two types, using two different algorithms. **xnu** masks repeated sequences while **seg** masks regions of low-complexity - regions where there are "too many" serines for example. Masking for low-complexity stops you hitting sequences that are similar to your query sequence only because they both have similar compositional bias: proline-rich proteins for example. An example follows:

```
>P04729 Wheat gamma gliadin
MKTFLVFALIAVVATSAIAQMETSCISGLERPWQQQPLPPQQSFSQQPPFSQQQQQPLPQ
QPSFSQQQPPFSQQQPILSQQPPFSQQQQPVLPQQSPFSQQQQLVLPPQQQQQQLVQQQI
PIVQPSVLQQLNPCKVFLQQQCSPVAMPQRLARSQMWQQSSCHVMQQQCCQQLQQIPEQS
RYEAIRAIIYSIILQEQQQGFVQPQQQQPQQSGQGVSQSQQQSQQQLGQCSFQQPQQQLG
QQPQQQQQQQVLQGTFLQPHQIAHLEAVTSIALRTLPTMCSVNVPLYSATTSVPFGVGTG
VGAY*
```

and after low complexity masking:

```
>P04729 SEG low-complexity masked
MKTFLVFALIAVVATSAIAQMETSCISGLERPWXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXLNPCKVFLQQQCSPVAMPQRLARSQMWXXXXXXXXXXXXXXXXXXXXXXX
RYEAIRAIIYSIIXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXHQIAHLEAVTSIALRTLPTMCSVNVPLYSATTSVPFGVGTG
VGAY*
```

Similar filtering (another word for masking) can be carried out on DNA sequences with a program called DUST. This will effectively erase such minimally informative but very widely distributed sequences as polyA tails.

Scoring matrices.

Homology searching algorithms all look for the best matches between the query sequence and database sequences. "best" is defined by a high score using one of several alternative scoring matrices. One such matrix - blosum62 - is shown below. This matrix is based on observed substitutions in a database of aligned sequences where 62% of the residues are identical. The distribution of the remaining 38% is analysed to yield:

```
# BLOSUM 62
A R N D C Q E G H I L K M F P S T W Y V
A 4 -1 -2 -2 0 -1 -1 0 -2 -1 -1 -1 -1 -2 -1 1 0 -3 -2 0
R -1 5 0 -2 -3 1 0 -2 0 -3 -2 2 -1 -3 -2 -1 -1 -3 -2 -3
N -2 0 6 1 -3 0 0 0 1 -3 -3 0 -2 -3 -2 1 0 -4 -2 -3
```

```
D -2 -2 1 6 -3 0 2 -1 -1 -3 -4 -1 -3 -3 -1 0 -1 -4 -3 -3
C 0 -3 -3 -3 9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1
Q -1 1 0 0 -3 5 2 -2 0 -3 -2 1 0 -3 -1 0 -1 -2 -1 -2
E -1 0 0 2 -4 2 5 -2 0 -3 -3 1 -2 -3 -1 0 -1 -3 -2 -2
G 0 -2 0 -1 -3 -2 -2 6 -2 -4 -4 -2 -3 -3 -2 0 -2 -2 -3 -3
H -2 0 1 -1 -3 0 0 -2 8 -3 -3 -1 -2 -1 -2 -1 -2 -2 2 -3
I -1 -3 -3 -3 -1 -3 -3 -4 -3 4 2 -3 1 0 -3 -2 -1 -3 -1 3
L -1 -2 -3 -4 -1 -2 -3 -4 -3 2 4 -2 2 0 -3 -2 -1 -2 -1 1
K -1 2 0 -1 -3 1 1 -2 -1 -3 -2 5 -1 -3 -1 0 -1 -3 -2 -2
M -1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5 0 -2 -1 -1 -1 -1 1
F -2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 0 6 -4 -2 -2 1 3 -1
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7 -1 -1 -4 -3 -2
S 1 -1 1 0 -1 0 0 0 -1 -2 -2 0 -1 -2 -1 4 1 -3 -2 -2
T 0 -1 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 1 5 -2 -2 0
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11 2 -3
Y -2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 -1 3 -3 -2 -2 2 7 -1
V 0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4
```

Exercise:
Use the matrix to verify that the following sequence match clipped from a blast homology search has the right score (the convention is that exact matches are echoed on the middle line, "mismatches" have nothing, while "conservative substitutions", such as the replacement of leucine by isoleucine below, are given a +):

```
Score = 28:
Query: 3 LKQSNTLL 10
L QSNT+L
Sbjct: 62 LYQSNTIL 69
```

Choosing a different scoring matrix will give you a different cohort of hits.

```
#BLOSUM 30
A R N D C Q E G H I L
A 4 -1 0 0 -3 1 0 0 -2 0 -1
R -1 8 -2 -1 -2 3 -1 -2 -1 -3 -2
N 0 -2 8 1 -1 -1 -1 0 -1 0 -2
D 0 -1 1 9 -3 -1 1 -1 -2 -4 -1
C -3 -2 -1 -3 17 -2 1 -4 -5 -2 0
Q 1 3 -1 -1 -2 8 2 -2 0 -2 -2
E 0 -1 -1 1 1 2 6 -2 0 -3 -1
G 0 -2 0 -1 -4 -2 -2 8 -3 -1 -2
H -2 -1 -1 -2 -5 0 0 -3 14 -2 -1
I 0 -3 0 -4 -2 -2 -3 -1 -2 6 2
L -1 -2 -2 -1 0 -2 -1 -2 -1 2 4

#BLOSUM 90
A R N D C Q E G H I L
A 5 -2 -2 -3 -1 -1 -1 0 -2 -2 -2
R -2 6 -1 -3 -5 1 -1 -3 0 -4 -3
N -2 -1 7 1 -4 0 -1 -1 0 -4 -4
D -3 -3 1 7 -5 -1 1 -2 -2 -5 -5
C -1 -5 -4 -5 9 -4 -6 -4 -5 -2 -2
Q -1 1 0 -1 -4 7 2 -3 1 -4 -3
E -1 -1 -1 1 -6 2 6 -3 -1 -4 -4
G 0 -3 -1 -2 -4 -3 -3 6 -3 -5 -5
H -2 0 0 -2 -5 1 -1 -3 8 -4 -4
I -2 -4 -4 -5 -2 -4 -4 -5 -4 5 1
```

```
L -2 -3 -4 -5 -2 -3 -4 -5 -4 1 5
```

Compare the scores of following two alignments using blosum30 and blosum90

```
Alignment Score Matrix Score Alignment
Query: GHDEICI 39 Blos30 19 Query: HEQCRLEN
GH + C +E LEN
Sbjct: GHACNCG 5 Blos90 24 Sbjct: QENAHLEN
```

In the examples above, Blosum 30 will give a higher score to and thus preferentially find the GHDEICI match while Blosum 90 will find HEQCRLEN. In real database searches changing the substitution matrix may change the order in which sequences are scored and reported, in other cases it will identify totally different sequences as having a relationship with the query sequence.

Expectation cutoff

The blast defaults are designed to suit most of the people most of the time. In order to minimise the collection of marginal, statistically non-significant information, blast sets an 'expectation cutoff' parameter to 10. Accepting this means that blast will not report any match so common that you would expect to find 10 copies in the database by chance alone. A search for a short protein motif, ELVIS for example, in Swissprot with its 77,000 entries and 2 million residues will, by chance alone, find several to many copies. If you are using blastp for such a short motif search then you should crank up the expectation cutoff to the maximum of 1000. On the other hand, if you are only interested in very precise homologues and do not wish to be overwhelmed with a flood of marginal alignments, you might consider setting the E value to 0.001

Limit search taxonomically

Most Blast servers now will allow you to choose a subset of the sequence universe to search against. You should be able to search only human sequences or only mammalian sequences for example.

Output delivery options.

While blast is a general workhorse for finding similar sequences, each researcher will be asking a more or less specific question of their search. If you want to see if your sequence is homologous with anything, then a single hit would be enough. If you wanted to find all members of a protein family, perhaps to align them to find conserved residues, then more then 200 hits might not be enough. The quantity of information returned by a typical blast search can be substantial and will consume large amounts of disk to store it and many trees to print it. Accordingly, you are given the option to limit a) the number of hits and b) the number of alignments reported. Good servers will give you the option of returning the output in HTML with clickable links to the relevant database entries.

WWW access to Blast.

You can access blast in many different ways at many different sites. These are NOT all equivalent! The default parameters may be significantly different, the databases may not be updated on the same schedule and so may be significantly different in size or level of redundancy. Three accessible, authoritative, alternatives are on the WWW.

The Blast server at the NCBI in Bethesda, MD, USA:

http://www.ncbi.nlm.nih.gov/BLAST

The Blast server at the EBI in Hinxton, UK:

http://www.ebi.ac.uk/searches/searches.html

The SIB blast site is easily customizable

http://www.ch.embnet.org/

ILRI-BecA blast site

http://hpc.ilri.cgiar.org/bwb/

Blast guidelines.

When to use what algorithm

a. As a rule of thumb, if your DNA sequence is coding (i.e. not an intron, a structural RNA, "junk" DNA or some upstream control region), you should translate it first and use **blastp** search a protein database. It will be quicker, more sensitive and find more distant relatives.

b. If your DNA sequence is not coding, use **Fasta** instead. You should, therefore, rarely have to use blastn.

c. If you want to do a preliminary check for frameshift errors in your sequence, use **blastx** to compare your sequence, translated in all six reading frames, against a protein database. Why might this help you identify frameshift errors?

d. If you want to search for a particular protein sequence in a database of expressed sequence tags (ESTs) you will have to use **tblastn**.

A widely applicable blast protocol

If you want to carry out a reasonably comprehensive search of a protein database to find potential homologues to a query sequence you will have to carry out several

blastp searches. You will however, adjust your approach depending on the exact type of information that will satisfy your quest. On any well designed blast server it should be easy to determine what are the available options, but you should scrutinise the page carefully to determine what are the default options and parameters. By all means take the defaults, but, on its own this is unlikely to result in an adequate, let alone comprehensive, search. The DNA databases are doubling in size every 12-14 months; so a fresh blast search just before submitting your paper has much to recommend it.

On any reputable WWW homology server:

a. Paste in your sequence and do a search taking the default parameters.
b. Do the search again, with or without low-complexity masking, depending on what option the server has chosen as the default in part a. If low complexity regions are found the XXXed sequence should appear at the top of your results.
c. Do the search again using two different substitution scoring matrices. One based on sequences that are evolutionarily "close" such as Blosum90 or PAM30 and another based on sequences that are evolutionarily "distant" such as Blosum40 or PAM250. The latter search is more likely to pick up a rather distant, diffuse weak homologue.
d. If appropriate (sometimes your sequence will have no low-complexity regions) do b x c to carry out, in all, six blast searches.
e. If your results indicate that the first 100s of best hits are members of a well characterised protein family (a fact that you may already know), and that these hits are all pointing to a particular domain of your query protein, you may have to edit (by hand!) your sequence (XXXXing out the already identified region) to find more distant and potentially interesting homologues which have been swamped out by a deluge of higher scoring hits.
f. Scrutinise the results of all your searches taking into account not only the scores but also the alignments. Pay particular attention to hits which are unexpected or counter-intuitive.
g. You can eliminate a large number of useless but positive hits by only searching, say, human sequences.

<u>Interpreting output from blastp.</u>

Output from a blast search is voluminous and in four or five parts.

1. The first part is administrative, and should include copyright information, the date, references and most importantly a note of what database has been searched and what size it was. With the DNA database doubling in size every year, you will not be able to 'replicate your blast experiment' after an interval of as little as two weeks. You should note down these details for your materials and methods section.
2. On some sites (NCBI) a very useful graphic showing the length and degree of homology of all the hits follows. You can ‘mouse-over’ this to see which sequences are homologous to (part of) your query.

3. There follows a list of "hits" with a) a database accession number or other identifier b) a brief description c) a score and d) some information on the probability of finding such a hit in the searched database. There will be a certain amount of variation among servers in how this information is presented.
4. After this there are a number of alignments of the query sequence with the significant hits.
5. Finally there is more administrative and statistical information including any warnings or error messages.

The hit list should look like:

Blast server EBI:

```
Sequences producing significant alignments: (bits) Value
SW:GDB1_WHEAT P04729 GAMMA-GLIADIN B-I PRECURSOR. 616 e-176
SW:GLTC_WHEAT P16315 GLUTENIN, LOW MOLECULAR WEIGHT SUBUNIT ... 510 e-144
SW:GLTB_WHEAT P10386 GLUTENIN, LOW MOLECULAR WEIGHT SUBUNIT ... 480 e-135
SW:GLTA_WHEAT P10385 GLUTENIN, LOW MOLECULAR WEIGHT SUBUNIT ... 343 3e-94
SW:GDB3_WHEAT P04730 GAMMA-GLIADIN (GLIADIN B-III) (FRAGMENT). 329 5e-90
SW:HOR1_HORVU P06470 B1-HORDEIN PRECURSOR. 323 3e-88
SW:HOR3_HORVU P06471 B3-HORDEIN (FRAGMENT). 310 3e-84
```

Then after a large number of 'sensible' hits, such reports as:

```
SW:INVO_RAT P48998 INVOLUCRIN. 61 4e-09
SW:SRY_MOUSE Q05738 SEX-DETERMINING REGION Y PROTEIN (TESTIS... 61 4e-09
SW:FTSK_ECOLI P46889 CELL DIVISION PROTEIN FTSK. 59 2e-08
SW:OVO_DROME P51521 OVO PROTEIN (SHAVEN BABY PROTEIN). 58 2e-08
SW:FCA_ARATH O04425 FLOWERING TIME CONTROL PROTEIN FCA. 57 7e-08
SW:CLOC_MOUSE O08785 CIRCADIAN LOCOMOTER OUTPUT CYCLES KAPUT... 56 1e-07
SW:E75B_DROME P17672 ECDYSONE-INDUCIBLE PROTEIN E75-B. 52 1e-06
```

The 1e-06 on the last line of the output tells you that the probability of finding a match as good as this by chance in the current database is 1 * e-06. For biologists who are used to accepting probabilities of 0.05 or 0.001 as meaningful, this is highly significant statistically, but may nevertheless mean little or nothing biologically.
The first three hits are the same when you use the blast server at the NCBI but, because the implementation is different the probabilities are different. You'll have to be careful to record where, when and using what parameters you do your blast searches if you want them to be reproducible.
Blast server NCBI:

```
Sequences producing significant alignments: (bits) Value
gi|121100|sp|P04729|GDB1_WHEAT GAMMA-GLIADIN B-I PRECURSOR ... 197 2e-50
gi|121459|sp|P16315|GLTC_WHEAT GLUTENIN, LOW MOLECULAR WEIG... 176 3e-44
gi|121102|sp|P04730|GDB3_WHEAT GAMMA-GLIADIN (GLIADIN B-III... 114 2e-25
gi|123458|sp|P06470|HOR1_HORVU B1-HORDEIN PRECURSOR >gi|100... 103 4e-22
```

To make an estimate of the biological significance, you will have to look further down the output until you come to a listing of the alignments and scores of which the "hit-list" is a summary:

```
>SW:DC11_DROME P18169 drosophila melanogaster (fruit fly). defective chorion-1
fc125 protein precursor. 2/91 Length = 1123
Score = 215 (80.7 bits), Expect = 7.7e-16, P = 7.7e-16
Identities = 73/233 (31%), Positives = 119/233 (51%)
Query: 34 QQQPLPPQQ-SFSQQPPFSQQQQQPLPQQPSFSQQQPPFSQQQPILSQQPPFSQQQQPVL 92
```

```
QQ P+ QQ +S++ QQ QQ + Q P QQ+ +S++Q + QQ QQ P++
Sbjct:570 QQNPMMMQQRQWSEEQAKIQQNQQQIQQNPMMVQQRQ-WSEEQAKI-QQNQQQIQQNPMM 627
...
Query:149 QRLARSQMWQQSSCHVMQQQCCQQLQQIPEQSRYEAIRAIIYSIILQEQQQGFVQPQQQQ 208
Q R W + ++QQ QQ Q +Q+R + + + ++Q+Q+Q PQ Q
Sbjct:688 QMQQRQ--WTEDP-QMVQQM--QQRQWAEDQTRMQMAQQ---NPMMQQQRQMAENPQMMQ 739
Query:209 PQQSGQG---VSQSQQQSQQQLGQCSFQQPQQQLGQQPQ---QQQQQQVLQGT 255
+Q + + Q+QQ +QQ Q QQ QQ+ + Q QQQQ+Q++Q T
Sbjct:740 QRQWSEEQTKIEQAQQMAQQN--QMMMQQMQQRQWSEDQAQIQQQQRQMMQQT 790
```

You can see that almost all the matched residues are Q = Glutamine. It is doubtful if this means anything more than that both genes happen to have a lot of CAG and CAA codons! Certainly you'd want other independent information before concluding that Wheat Gamma Gliadin and this Drosophila gene share a recent common ancestor or a similar structure.

From the NCBI server, using low complexity masking, you find, among many other hits, the following alignment:

```
sp|P06471|HOR3_HORVU B3-HORDEIN
Length = 264
Score = 62.5 bits (149), Expect = 1e-09
Identities = 32/63 (50%), Positives = 38/63 (59%)
Query: 131 LNPCKVFLQQQCSPVAMPQRLARSQMWXXXXXXXXXXXXXXXXXXXXXXXXRYEAIRAIIY 190
LNPCKVFLQQQCSP+AM QR+ARSQM R+EA+RAI+Y
Sbjct: 111 LNPCKVFLQQQCSPLAMSQRIARSQMLQQSSCHVLQQQCCQQLPQIPEQLRHEAVRAIVY 170
Query: 191 SII 193
SI+
Sbjct: 171 SIV 173
```

This is meaningful both statistically and biologically because it turns out the hordein is a barley storage protein functionally equivalent to wheat gliadin.

Exercise:

1. Use SRS to find a mouse sequence in SwissProt. Try using any of the search criteria used on day 1 or try the one of the following keywords from the course homepage.

2. Carry out a blast search taking the default parameters to see if you can find a human or a yeast homologue. Try changing the substitution matrix or low complexity masking to see if you can alter the order or composition of the 'hits'.

NB. Do NOT submit another search until the first result is returned – especially at NCBI

Multiple Sequence Alignment

TOPICS

1. Introduction to multiple sequence alignment (MSA).
2. ClustalW
3. T-Coffee.
4. Multiple Sequence Alignment Editors.

Introduction

It is truism to say that there would be no genetics, and no very interesting biology, but for the fact that there is variability between individuals and among species. For years biological research depended on observable (bristle count, leaf size, plumage colour, colony morphology) variations. Then it became possible to document differences by using biochemical and other techniques (gram stain, lactose metabolism, blood groups). Over the last two or three decades it has become possible to get a rather direct measure of similarities and differences in the living world as molecular biologists have succeeded in cloning and sequencing DNA from an enormous variety of organisms. Notably a number of complete genomes have been completely sequenced over the last eight or so years, ultimately giving us the genetic and developmental blueprint for several living organisms. It is still many years before we will collectively be able to make complete sense of, say, the 4 million base pairs of the E. coli genome. Let alone the 1000x bigger human genome. One tool we have already used for making sense of sequence is homology searching.

Another widely used bioinformatic technique is to try to align several related sequences to find which residues/bases are conserved and which are variable. This will help in the understanding of the constraints under which the sequences may labour: conserved residues may be an essential part of the active site of an enzyme, variable residues may be part of a 'generic' alpha-helix.

Multiple sequence alignment is also a vital prerequisite for trying to determine the phylogenetic relationships among a group of related sequences - and by extrapolation between the species or varieties that contain those sequences.

Multiple sequence alignment is very computationally intensive. The numbers involved in evaluating all possible alignments between two sequences allowing gaps in either is large. When 3 or more sequences are involved the numbers become so large that the problem becomes incomputable. It requires an insight and a shortcut to get biologically informative alignments in a finite time. One of the earliest successful programs that could calculate a non-trivial multiple sequence alignment in a reasonable time was invented in TCD in 1986 by Des Higgins. We will be using web-based derivatives of the original clustal program that was written all those years ago for incredibly primitive pre-windows PCs. The program is also freely and widely available for PCs, Macs and Unix workstations. These standalone versions are probably more sensitive and convenient than the WWW based version.

ClustalW

http://www.ebi.ac.uk/clustalw/

http://www.ch.embnet.org/software/ClustalW.html

This web page allows you to make a multiple sequence alignment of any group (N >= 2!) of sequences. The poor thing will attempt to align whatever sequences you give it, but this may take a long time if the sequences are unrelated or numerous. This is another example where a user-friendly program, which makes a lot of choices for you by default, can be a poisoned chalice. There is a tendency among users to believe that the computer or the program does the alignment and that this excuses the humans involved from exercising judgment. There is even a widespread belief that changing the options or particularly editing a delivered alignment is somehow "unscientific" because it requires a subjective assessment of what is correct, sensible and meaningful. This wrong-headed attitude is frequently compounded by loading the computer generated multiple sequence alignment directly into a phylogenetic tree drawing algorithm to determine the relationships amongst the included taxa. Such a phylogeny program will, like ClustalW, try to do what it is asked to do and may generate a tree that is, shall we say, fatuous.

Clustal is a program for computer-**aided** multiple sequence alignment. It takes some of the grunt work out of the complex and time-consuming business of aligning many sequences. It does this by the judicious insertion of gaps to represent the insertions and deletions that have occurred over evolutionary time since the most recent common ancestor of the sequences included. All users of the program are morally and scientifically obliged to scrutinize critically the alignment and see how it can be improved. There are numerous, colorful, multiple sequence alignment editors available to help you do this.

The ClustalW home page is nicely designed because all the options and parameters are visible on the one page as choice buttons. You can get a little help on the effect of each of these choices by clicking on the hypertext link above the choice button. Rather more information on the theory and practice of Clustal can be found at:

http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html

The Clustal WWW servers invites you to "Enter or Paste a set of Sequences in any Format", an invitation which should be treated with caution. FASTA format has much to recommend it. In this format, each sequence is represented by a single title line beginning with a ">" followed by the sequence itself on subsequent lines; typically 60 residues or bases per line, thus:

```
>ACDRECAP.RECA 355
MDEPGGKIEFSPAFMQIEGQFGKGAVMRAGDKPGINDPDVKSTGSLGLDGALGQGGLPRG
RVVEIYGPESSGKTTLTLKAIASAQAEGATPAFTDAEHALDPGFASKLGVNVKRLLISQP
DTGEQALEIADMLFRSGAVDVIVKDSVAALTPKAEIEGEMGDSHQGLHARLMSQALRNKT
ANISRWNKLVIFKKQIRMKMGVYGRPETTTGGNALKFYASVRLDIRRMGAMKKSATKSYD
WSTRVKVVKNKVAPPFRQAELAIYYGEGIYRGSEPVDLGVKLENVEKSGGWYSYPGRRIG
QGKANARQYLRVKPEFPGIFEQGIRGAMAAPHPLGFGERRDVQQESGEPYGNNGX
>BRURECA.RECA 361
MSQNSLRLVEDNSVDKTKALDAALSQIERAFGKGSIMRLGQNDQVVEIETVSTGSLSLDI
ALGVGGLPKGRIVEIYGPESSGKTTLALHTIAEAQKKGGICAFVDAEHALDPVYARKLGV
HLENLLISQPITGEQALEITDTLVRSGAIDVLVVDSVAALTPRAEIEGEMGDSHGLQARL
MSQAVRKLTGSISRSNCMVIFINQIRMKIGVMFGSPETTTGGNALKFYASVRLDIRRIGS
IKERDEVVGNQTRVKVVKNKLAPPFKQVEFDIMYGAGVSKVGELVDLGVKAGVVEKSGAW
FSYNSQRLGQGRENAKQYLKDNPEVAREIETTLRQNAGLIAEQFLDDGGPEEDAAGAAMX
>NGRECAG.RECA 349
MSDDKSKALAAALAQIEKSFGKGAIMKMDGSQQEENLEVISTGSLGLDLALGVGGLRRGR
IVEIFGPESSGKTTLCLEAVAQCQKNGGVCAFVDAEHAFDPVYARKLGVKVEELYLSQPD
TGEQALEICDTLVRSGGIDMVVVDSVAALVPKAEIEGDMGDSHVGLQARLMSQALRKLTG
HIKKTNTLVVFINQIRMKIGVMFGSPETTTGGNALKFYSSVRLDIRRTGSIKKGEEVLGN
ETRVKVIKNKVAPPFRQAEFDILYGEGISWEGELIDIGVKNDIINKSGAWYSYNGAKIGQ
GKDNVRVWLKENPEISDEIDAKIRALNGVEMHITEGTQDETDGERPEEX
```

With a very highly conserved protein (histones or mammalian beta globins or recA from gamma proteobacteria) it may well be possible to align sequences by hand and eye and good judgement, using, say, Microsoft WORD. Nevertheless, this is likely to be a time consuming process and becomes impossible if many gaps are required or if the evolutionary relationship between the sequences is more tenuous.

Clustal works in a three step-process:

1) All sequences are aligned and compared to each other and a score or 'distance' is calculated between each pair of sequences.

2) This matrix of distances between each pair of sequences is used to create a 'dendrogram' or phylogenetic tree among the included sequences. (This was Des Higgins' key insight that cracked the problem open)

3) The dendrogram is used as the basis for constructing the real multiple sequence alignment: basically the most closely related sequences or groups of sequences are aligned first.

The quality of the alignment is determined by assigning a positive score to each pair of identical residues which is aligned, and a lower or negative score to 'mismatches'.

The scores are read off from the substitution matrix which is in force (by default or by choice). See Chapter on BLAST for more on substitution matrices.

The parameters most likely to affect the quality of the alignment are the gap penalty (GAP OPEN), the gap-extension penalty (GAP EXTENSION) and, to a lesser extent, the substitution matrix (MATRIX).

Gap Open Penalty.

If you attempt to align two sequences starting at the amino terminus or the 5' end of the sequences and one of the sequences has a deletion, then the alignment is likely to be very poor after the deletion unless a gap is inserted. This gap mimics the biological reality that one sequence has lost one or more residues/bases. Usually we don't know where the deletion has occurred or indeed if it is really an insertion in the other sequence. Clustal attempts to estimate where such a deletion is most likely to have happened. It does this with a Gap Penalty. The gap penalty is typically more negative than the 'worst' mismatch. If the gap is correctly sited then the negative score incurred by the gap penalty will be more than compensated for by enhanced positive scores further down the alignment. A high gap penalty will discourage gaps, while a very low gap penalty will allow gaps willy-nilly and so enable you to align two completely unrelated sequences.

Gap Extension Penalty.

Most sequence alignment programs that work well use what are called affine gap penalties, so that a gap of three bases/residues is not penalised three times more heavily that a gap of one. This is taking account of the fact that a point deletion is more or less as common as a longer one. So taking the default gap penalties from the clustalWWW server (Open = 10, Ext=0.05) we get a score of -10 for a single residue gap and -10.45 (10 + 9*0.05) for a gap of ten residues.

T-COFFEE

For distant or difficult alignments T-COFFEE is almost certain to give you a better result than clustalW. It is freely available for download but is also available over the web. http://www.ch.embnet.org/software/TCoffee.html

Paste your PROTEIN sequences into the box on this page and click on the [run T-COFFEE] box. When the run is finished a

```
Here are your search results:
```

Will appear. There are a number of formats for outputting your alignment. You are advised to choose **phylip** output if you plan to use that software suite for constructing phylogenetic trees.

Exercise:

1) Choose any 5-10 sequences from the same family (defined by prosite?) or from the results of a homologue search. Or from the list of mammalian sequences which have more than one representative in SwissProt which are on the course homepage.

2) Run them through the clustal WWW server taking the default parameters.

3) Critically evaluate the alignment:

a) if one sequence is much shorter than the others find out why - a partial sequence ?

b) if one or two sequences seem to be distorting the alignment, consider ejecting them and redoing the alignment.

c) can you improve the alignment by choosing different gap penalties ?

5) If you can get a good alignment use the Jpred Predict Protein prediction server at the EBI to see if the gaps appear in peptide loops (that might not be expected to be essential to the structure and function of the enzyme).

6) Can you find the prosite motif that defines your family of proteins in your multiple sequence alignment? Are the elements of that motif always conserved?

7) Does T-COFFEE make a better fist of a "difficult" multiple sequence alignment like the casein dataset?

Multiple sequence alignment editors

For reasons outlined at the beginning of this chapter it is important not to treat multiple sequence alignment software as a black-box. You must scrutinize the alignment created and almost certainly you will want to do some editing to align motifs, cysteines, and hydrophobic residues. Each alignment will be different and you can look up SwissProt or Pfam to discover structural information about, and conserved residues peculiar to, your protein (family) of interest. Obviously, T-Coffee and ClustalW can't read PubMed, SwissProt – that's your job. Try these MSA editors:

On the WWW:

JalView: http://www.jalview.org

For MS-Windows:

Genedoc: http://www.psc.edu/biomed/genedoc

Note: There are many others

Printed sources about BioInformatics & the Internet.

Briefings in Bioinformatics - a journal aimed at users rather than developers with useful review and how-to articles.

Books:

Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins. Andreas Baxevanis & B.F.Francis Ouellette (Eds). John Wiley & Sons 2nd Ed 2001; ISBN: 0471-38390-2 The Course text book!

Fundamentals of Molecular Evolution. W-H Li and D Graur. Sinauer 1991. ISBN 0 87893 452 9

Fundamentals of Molecular Evolution. D Graur and W-H Li . Sinauer 2000. ISBN 0-87893-266-6

PAUP 4.0 Phylogenetic Analysis Using Parsimony (and other methods) Manual. David L Swofford. Sinauer 1999. 0 87893 801 X

Introduction to Bioinformatics. TK Attwood & DJ Parry-Smith. Addison Wesley Longman 1999. ISBN 0582 32788 1.

Molecular Evolution: a phylogenetic approach. RDM Page and EC Holmes. Blackwell 1998. ISBN: 0-86542-889-1

Bioinformatics for Dummies. Notredame and Claverie. 2003

Articles:

Baldauf, SL (2003) Phylogeny for the faint of heart: a tutorial. TIG 19(6): 345-351.

APPENDIX I

SEQUENCE SYMBOLS

Nucleotides

IUBCODE	MEANING	COMPLEMENT
A	A	T
C	C	G
G	G	C
T/U	T	A
M	A OR C	K
R	A OR G	Y
W	A OR T	W
S	C OR G	S
Y	C OR T	R
K	G OR T	M
V	A OR C OR G	B
H	A OR C OR T	D
D	A OR G OR T	H
B	C OR G OR T	V
X/N	G OR A OR T OR C	X
.	NOT	G OR A OR T OR C

Amino Acids

SYMBOL	MEANING	CODONS	IUB CODE
A	ALA	GCT, GCC, GCA, GCG	!GCX
B	ASP, ASN	GAT, GAC, AAT, AAC	!RAY
C	CYS	TGT, TGC	!TGY
D	ASP	GAT, GAC	!GAY
E	GLU	GAA, GAG	!GAR
F	PHE	TTT, TTC	!TTY

G	**GLY**	**GGT, GGC, GGA, GGG**	**!GGX**
H	**HIS**	**CAT, CAC**	**!CAY**
I	**ILE**	**ATT, ATC, ATA**	**!ATH**
K	**LYS**	**AAA, AAG**	**!AAR**
L	**LEU**	**TTG, TTA, CTT, CTC, CTA, CTG**	**!TTR, CTX, YTR**
M	**MET**	**ATG**	**!ATG**
N	**ASN**	**AAT, AAC**	**!AAY**
P	**PRO**	**CCT, CCC, CCA, CCG**	**!CCX**
Q	**GLN**	**CAA, CAG**	**!CAR**
R	**ARG**	**CGT, CGC, CGA, CGG, AGA, AGG**	**!CGX, AGR, MGR**
S	**SER**	**TCT, TCC, TCA, TCG, AGT, AGC**	**!TCX, AGY**
T	**THR**	**ACT, ACC, ACA, ACG**	**!ACX**
V	**VAL**	**GTT, GTC, GTA, GTG**	**!GTX**
W	**TRP**	**TGG**	**!TGG**
X	**UNKNOWN**		**!XXX**
Y	**TYR**	**TAT, TAC**	**!TAY**
Z	**GLU, GLN**	**GAA, GAG, CAA, CAG**	**!SAR**
*****	**TERMINATOR**	**TAA, TAG, TGA**	**!TAR, TRA**